Convolutional neural networks to automate the screening of malaria in low-resource countries (#48772)

First submission

Guidance from your Editor

Please submit by **5 Jun 2020** for the benefit of the authors (and your $200 publishing discount) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

1 Latex file(s)

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Speculation is welcome, but should be identified as such.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Convolutional neural networks to automate the screening of malaria in low-resource countries

Oliver Zhao [Corresp., 1], **Nikhil Kolluri** [2], **Annie Anand** [1], **Nicholas Chu** [2], **Ravali Bhavaraju** [1], **Aditya Ojha** [2], **Sandhya Tiku** [1], **Dat Nguyen** [1], **Ryan Chen** [1], **Adriane Morales** [1], **Deepti Valliappan** [1], **Juhi P Patel** [3], **Kevin Nguyen** [3]

[1] Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, United States of America

[2] Department of Electrical & Computer Engineering, The University of Texas at Austin, Austin, TX, United States of America

[3] Department of Psychology, The University of Texas at Austin, Austin, TX, United States of America

Corresponding Author: Oliver Zhao
Email address: oliver.zhao@utexas.edu

Malaria is an infectious disease caused by Plasmodium parasites, transmitted through mosquito bites. Symptoms include fever, headache, and vomiting, and in severe cases, seizures and coma. The World Health Organization reports that there were 228 million cases and 405,000 deaths in 2018, with 93% and 94% of total malaria cases and deaths occurring in Africa, respectively. Rapid diagnosis and subsequent treatment is the most effective means to mitigate the progression into serious symptoms. However, many fatal cases have been attributed to poor access to healthcare resources for malaria screenings. In these low-resource settings, the use of light microscopy on a thin blood smear with Giemsa stain is used to examine the severity of infection, requiring tedious and manual counting by a trained technician.

To address the malaria endemic in Africa and its coexisting socioeconomic constraints, we propose an automated, mobile phone-based, screening process that takes advantage of already existing resources. Through the use of convolutional neural networks (CNNs), we utilize a SSD multibox object detection architecture that rapidly processes thin blood smears acquired via light microscopy to isolate images of individual red blood cells with 90.4% average precision. Then we implement a FSRCNN model that upscales 32x32 low-resolution images to 128x128 high-resolution images with a PSNR of 30.2, compared to a baseline PSNR of 24.2 through traditional bicubic interpolation. Lastly, we utilize a modified VGG16 CNN that classifies red blood cells as either infected or uninfected with an accuracy of 96.5% in a balanced class dataset. These sequential models create a streamlined screening platform, giving the healthcare provider the number of malaria-infected red blood cells in a given sample. Our deep learning platform is efficient enough to operate exclusively on low-tier smartphone hardware, eliminating the need for trained diagnostic technicians and high-speed internet connection.

Convolutional Neural Networks to Automate the Screening of Malaria in Low-Resource Countries

Oliver S. Zhao[1], Nikhil Kolluri[2], Annie Anand[1], Nicholas Chu[2], Ravali Bhavaraju[1], Aditya Ojha[2], Sandhya Tiku[1], Dat Nguyen[1], Ryan Chen[1], Adriane Morales[1], Deepti Valliappan[1], Juhi Patel[3], and Kevin Nguyen[3]

[1]Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, United States of America
[2]Department of Electrical Engineering, The University of Texas at Austin, Austin, TX, United States of America
[3]Department of Psychology, The University of Texas at Austin, Austin, TX, United States of America

Corresponding author:

Oliver S. Zhao[1]

Email address: oliver.zhao@utexas.edu

ABSTRACT

Malaria is an infectious disease caused by *Plasmodium* parasites, transmitted through mosquito bites. Symptoms include fever, headache, and vomiting, and in severe cases, seizures and coma. The World Health Organization reports that there were 228 million cases and 405,000 deaths in 2018, with 93% and 94% of total malaria cases and deaths occurring in Africa, respectively. Rapid diagnosis and subsequent treatment is the most effective means to mitigate the progression into serious symptoms. However, many fatal cases have been attributed to poor access to healthcare resources for malaria screenings. In these low-resource settings, the use of light microscopy on a thin blood smear with Giemsa stain is used to examine the severity of infection, requiring tedious and manual counting by a trained technician.

To address the malaria endemic in Africa and its coexisting socioeconomic constraints, we propose an automated, mobile phone-based, screening process that takes advantage of already existing resources. Through the use of convolutional neural networks (CNNs), we utilize a SSD multibox object detection architecture that rapidly processes thin blood smears acquired via light microscopy to isolate images of individual red blood cells with 90.4% average precision. Then we implement a FSRCNN model that upscales 32x32 low-resolution images to 128x128 high-resolution images with a PSNR of 30.2, compared to a baseline PSNR of 24.2 through traditional bicubic interpolation. Lastly, we utilize a modified VGG16 CNN that classifies red blood cells as either parasitized or uninfected with an accuracy of 96.5% in a balanced class dataset. These sequential models create a streamlined screening platform, giving the healthcare provider the number of malaria-infected red blood cells in a given sample. Our deep learning platform is efficient enough to operate exclusively on low-tier smartphone hardware, eliminating the need for trained diagnostic technicians and high-speed internet connection.

INTRODUCTION

Malaria in Developing Countries

Malaria is an infectious disease caused by *Plasmodium* parasites, which are transmitted through female mosquito bites. *P. falciparum* is the most common and the deadliest human malaria parasite in Africa, accounting for nearly all fatal cases in Sub-Saharan Africa (WHO, 2019), (F. Ellis McKenzie, 2008), (Rasheed O. Makanjuola, 2020). Typical symptoms include fever, malaise, headaches, and vomiting, and in severe cases, seizures and coma. The World Health Organization (WHO) reports that in 2018, there were 228 million cases and 405,000 deaths globally. 93 and 94 percent of total malaria cases and deaths occurred in Africa, respectively (WHO, 2019). The most vulnerable group of infected individuals are children under the age of five, where 67% of malaria deaths occur. The WHO suggests that rapid diagnosis and subsequent treatment is the most effective means to mitigate the progression into serious symptoms. However, less than 29% of children under the age of five in sub-Saharan Africa receive antimalarial drug treatment (WHO, 2019), despite this demographic being at the greatest risk (Ricci, 2012). The WHO cites that significant factors driving this statistic are poor access to healthcare and ignorance of malaria symptoms (WHO, 2019).

Malaria can be diagnosed based on clinical symptoms, although the Center for Disease Control (CDC) always recommends confirming the diagnosis with a laboratory test (CDC, 2020). Laboratory tests can include the use of PCR to identify the specific strain of *Plasmodium* in a confirmed malaria case (Nguyen Van Hong and Erhart, 2013), antigen detection kits to detect *Plasmodium*-derived antigens (Duangporn Polpanich, 2007),(Haris M Khan, 2010), and serology tests such as ELISA to detect antibodies targeting malaria parasites (Linda M. Murungi, 2019). These methods are expensive and often infeasible to implement in low-resource settings due to the required equipment and use of trained technicians (CDC, 2020). In low-resource settings, the use of light microscopy on a thin or thick blood smear with Giemsa stain is often used to confirm the presence of malaria parasites (E. Charpentier, 2020). However, the diagnostic accuracy of using Giemsa-strained thin blood smears depends heavily on the level of expertise in the technician, who must manually classify and count the number of malaria-infected red blood cells. This results in significant inter-observer variability due to the different levels of expertise in technicians in low-resource settings, who often have to learn other tasks and cannot be adequately trained for this specific task as a result (Mounkaila Abdou Billo, 2013),(Katherine M. Bowers, 2009).

Use of Machine Learning in Clinical Applications and Malaria Screening

The use of machine learning methods, particularly neural networks, is rapidly growing in areas of clinical application. The two primary applications are involved with either segmentation or classification in clinical images (Dinggang Shen, 2017), (Syied Anwar, 2018), (Geert Litjens, 2017) or histological images (Kan, 2017), (Shidan Wang, 2019). In particular, the use of machine learning to diagnose malaria is of interest, where various classification models are developed by several groups to determine whether a red blood cell is infected or uninfected, as shown in Table 1.

To address the severe malaria endemic in Africa and its related issues with medical resources and clinical expertise, we propose a multi-step automated screening process that takes advantage of readily available resources in low-income settings. Through the use of convolutional neural networks (CNNs), we utilize a SSD300 multibox model for object detection that rapidly processes Giemsa-stained thin blood smears acquired from basic light microscopy in order isolate images of individual red blood cells. Then we implement a separate FSRCNN image resolution upscaling model to raise the low resolution images of 32x32 pixels to 128x128 pixels, if necessary. Lastly, we utilize a variant of a VGG16 CNN that classifies every red blood cell as either infected or uninfected. These sequential models serve to create a streamlined mechanism from which our screening platform takes in thin blood smear images as inputs to provide the healthcare provider with the number and percentage of malaria-infected red blood cells in a given sample. Taking advantage of the prevalent availability of low-end smartphones in the African continent, our deep learning platform is lean and efficient enough to operate exclusively on the smartphone hardware, eliminating the need for high-speed internet access to transmit image information into a cloud-based neural network model.

Source	Accuracy	Sensitivity	Specificity	Dataset
(Nicholas E. Ross, 2006)	73.0	85.0	NR	Private
(Dev Kumar Das, 2013)	93.24	94.04	87.93	Private
(Kusworo Adi, 2016)	87.14	NR	NR	Private
(Zhaohui Liang, 2017)	97.37	96.99	97.75	NIH
(Yuhang Dong, 2017)	98.1	97.29	98.69	Private
(Kristofer E. Delas Peñas, 2017)	92.4	95.2	84.7	Private
(Gopalakrishna Pillai Gopakumar, 2017)	97.7	NR	NR	Private
(Sivaramakrishnan Rajaraman, 2018)	98.6	98.1	99.2	NIH
(Aimon Rahman, 2019)	97.71	97.48	97.94	NIH
(Sivaramakrishnan Rajaraman, 2019)	99.51	NR	NR	NIH

Table 1. Previous attempts by other research groups to classify infected red blood cells. A significant number of the groups used their own datasets, while other groups used the NIH dataset. NR = not reported.

METHODS

Dataset and Computing Platform

Two datasets from different sources were used: (1) NIH malaria dataset and (2) Broad Institute malaria dataset. The publicly available NIH malaria dataset was acquired from the Lister Hill National Center for Biomedical Communications (LHNCBC) at the National Library of Medicine (NLM), which contains 27,588 labeled and segmented cell images acquired from Giemsa-stained thin blood smear slides. The dataset contains equal instances of healthy red blood cells and *P. falciparium*-infected red blood cells derived from 150 *P. falciparium*-infected individuals and 50 healthy individuals. Meanwhile, the Broad Institute dataset contains 1364 blood smear images with 80,000 individually labeled blood cells that are either healthy or infected with *P. vivax*. In the Broad Institute dataset, only about 5% of the red blood cells are infected.

The Google Cloud Platform was utilized for acquiring the bulk of experimental data from training different variations of the neural network models. One of two machine configurations were used: (1) N1 machine with 8 vCPU and 52 GB memory with 1 Nvidia Tesla V100 GPU or (2) N1 machine with 16 vCPU with 104 GB memory and 2 Nvidia Tesla V100 GPUs. A boot disk with a Deep Learning on Linux operating system with the GPU Optimized Debian m32 (with CUDA 10.0) version was used. In addition, the free online Google Colab interface with a T4 GPU was used for rapid code write-up and subsequent preliminary testing.

Neural Network Performance Metrics

In all neural network models used for classification and resolution enhancement, five-fold cross-validation was performed in order to report the mean and standard deviation of the model performance. The cross-validation groups were randomly split and distributed evenly among the five groups, with the same set of cross-validation groups used to test different model variants in a given experiment. Positive and negative samples were defined as infected and uninfected red blood cells, respectively. Some experiments did not utilize the full dataset, instead using a randomly selected subset of the dataset to reduce computational burden.

The object detection model performance was measured through average precision and average recall across different conditions, such as the intersection over union (IoU) values, image sizes, and maximum number of detections. The following metrics were measured in the malaria classification model: accuracy, sensitivity, specificity, area under the curve (AUC), F1-score, and Matthews correlation coefficient (MCC). The image upscaling model measured the mean squared error (MSE) and peak signal-to-noise ratio (PSNR) to examine the quality of the image upscaling output. Bicubic interpolation was used as the baseline for measuring comparing the performance of the CNN-based resolution upscaling model. The

training and testing code and results are publicly available on a Github repository.

Development of Object Detection Model

A 300x300 Single Shot MultiBox Detector (SSD300) was trained to detect both infected and uninfected red blood cells from the thin blood smear images in the Broad Institute dataset. Because each red blood cell will be classified by the VGG16 classification model in later steps, the object detection model was not trained to distinguish between the two cell classes. The object detection model served primarily as a proof-of-concept to show that the mobile platform can sequentially run the object detection, resolution enhancement, and cell classification models in tandem. Consequently, the SSD300 model was not heavily fine-tuned to maximize performance. The final SSD300 model was trained with an RMSProp optimizer with a learning rate of 0.004. The batch size was 24 and the training process was run for 60,000 steps. All input images were scaled to the required 300x300 image size before entering the object detection model. The outputted thresholds from the 300x300 images were then rescaled to provide the original box coordinates of each individual red blood cell to isolate cropped images of each individual red blood cell.

Development of the Image Classification CNN

All input images from the NIH dataset were scaled to 128x128 resolution. In order to expand the number of hyperparameters examined, the CNN model was developed through sequential hyperparameter tuning rather than a traditional grid or random search. First, the feature extraction architecture was optimized before developing the classification architecture. Then, hyperparameters involved with the training of the model - such as the optimizer, learning rate, and batch size - were fine-tuned to give the final model. All experiments with the image classification CNN were performed on a subset of 10,000 randomly selected images to reduce computational burden. After the final classification CNN has been developed, the optimized hyperparameters were used to train on the entire dataset of 27,558 images to provide an accurate representation of the model performance.

Fine-Tuning the Feature Extraction Architecture During the fine-tuning of the feature extraction architecture, the following conditions were maintained for all experiments: (1) feature extraction layers were succeeded with two fully connected dense layers containing 512 nodes each with ReLU activation functions and 50% dropout, and (2) an Adam optimizer with a learning rate of 10^{-6} and batch size of 64 was used. The following pre-trained CNN architectures with weights initialize from the ImageNet dataset were used: ResNet50V2, VGG16, VGG19, InceptionV2, Xception, InceptionResNetV2, DenseNet121, and MobileNetV2. VGG16 and VGG19 are traditional deep CNNs (Karen Simonyan, 2015), while other models use residual connections to allow for deeper convolution layers (Kaiming He, 2016). It is also worthwhile to note that MobileNetV2 is designed specifically for mobile phone use, sacrificing accuracy for the sake of speed. The top performing model was chosen based on its overall accuracy and AUC. In the event of having similarly performing models, the model with the fewest parameters was selected to maximize model efficiency.

Fine-Tuning the Classification Architecture The number of nodes in each of the two fully connected dense layers were tested with 128, 256, 512, and 1024 nodes each, with the set of dense nodes that resulted in the highest accuracy and convergence speed chosen. Then, the following dropout rates were examined: 25%, 50%, and 75%. The dropout rate resulting in the highest convergence speed and lowest testing loss was chosen. Lastly, the rectified linear unit (ReLU) and Tanh activation functions were examined. When the given hyperparameter has yet to be fine-tuned, the experiments contained the following conditions: (1) 512 nodes in both dense layers, (2) 50% dropout, and (3) ReLU activation functions.

Optimizing the Learning Conditions The following optimizers were examined: Stochastic gradient descent (SGD) with Nesterov momentum, Adam, RMSProp, AdaMax, and Nadam. The following learning rates were tested: 10^{-6}, 10^{-5}, 10^{-4}, and 10^{-3}. Graphical results have not been shown for learning rates that failed to train the model, although tabular results are available on the Github repository. The optimal learning rates were selected from each optimizer. Then, the performances of each optimizer

were compared with the best optimizer chosen on the following three criteria: (1) final testing accuracy, (2) final testing loss, and (3) rate of convergence.

Development of CNN-Based Image Resolution Upscaler

The FSRCNN model was developed in 2016 as an improvement over the previous SRCNN model introduced in 2014 (Chao Dong, 2016), (Chao Dong, 2014). In short, the FSRCNN model performs feature extraction and shrinking a high dimensional feature map into a low dimensional feature map. Then a series of mapping layers are performed before the low dimensional feature map is expanded back to the high dimensional feature map. Finally, deconvolution is performed to generate the high resolution images. Consequently, the three main hyperparameters are: (1) number of mapping layers, (2) the dimension of the high feature map, and (3) the dimension of the low feature map. Consequently, we test using 2-4 mapping layers, 48 or 56 filters for high dimensional features, and 12 or 16 filters for low dimensional features.

In addition, we create two separate train and test sets to evaluate the effectiveness of the FSRCNN model: (1) FSRCNN-derived high resolution train and test sets and (2) bicubic interpolated high resolution train and test sets. These train and test sets are then used to train and validate the final malaria classification model to examine how the differences in image quality impacts the effectiveness of the classification CNN. Five-fold cross-validation with the full NIH dataset was used in these evaluations.

Implementation of TensorFlow Lite Android Platform

TensorFlow Lite is an open-source platform focused on on-device model inference. Unlike previously reported studies that utilize phone apps for model prediction (Sivaramakrishnan Rajaraman, 2019), this allows the models to run directly on the Android-based smartphones rather than relying on cloud-based computing resources. While all models were developed and trained with the TensorFlow and Keras packages, the final model deployments are subsequently converted into a .tflite file that allows the models to be run on the TensorFlow Lite package.

RESULTS

Red Blood Cell Object Detection Model

The SSD300 object detection model is able to detect the presence of red blood cells with an average precision of 90.4% when the IoU is 0.50 for all area sizes with 100 maximum detections, while the avearge recall is 63.9% at an IoU of 0.50:0.95 for all area sizes with 100 maximum detections, as shown in table 2. We see that the model has high precision, but relatively poor recall. In figure 1 we see an example of the bounding boxes and confidence levels of detected red blood cells from a sample image from the Broad Institute dataset.

Malaria Classification Model

Evaluating Pre-Trained Neural Network Architectures

Both the pre-trained neural network VGG16 and VGG19 architectures performed the best, both achieving approximately 0.9600 accuracy and an AUC of at least 0.9900, as shown in Table 3 and Figure 2. However, we see that VGG16 was slightly less prone to overfitting than VGG19, despite the slightly slower decline in testing loss. In addition, VGG16 requires slightly less operations to fit a slightly smaller amount of parameters. Consequently, the VGG16 model was selected for further hyperparameter tuning.

Metric Type	IoU	Area Size	Maximum Detections	Performance
Average Precision (AP)	0.50:0.95	all	100	AP = 0.436
Average Precision (AP)	**0.50**	**all**	**100**	**AP = 0.904**
Average Precision (AP)	0.75	all	100	AP = 0.491
Average Precision (AP)	0.50:0.95	small	100	AP = -1.00
Average Precision (AP)	0.50:0.95	medium	100	AP = 0.082
Average Precision (AP)	0.50:0.95	large	100	AP = 0.440
Average Recall (AR)	0.50:0.95	all	1	AR = 0.114
Average Recall (AR)	0.50:0.95	all	10	AR = 0.295
Average Recall (AR)	**0.50:0.95**	**all**	**100**	**AR = 0.639**
Average Recall (AR)	0.50:0.95	small	100	AR = -1.00
Average Recall (AR)	0.50:0.95	medium	100	AR = 0.144
Average Recall (AR)	0.50:0.95	large	100	AR = 0.605

Table 2. SSD300 performance metrics. Average precision and average recall across different IoUs, area sizes, and maximum number of detections. Top performing conditions for maximizing average precision and recall and bolded.

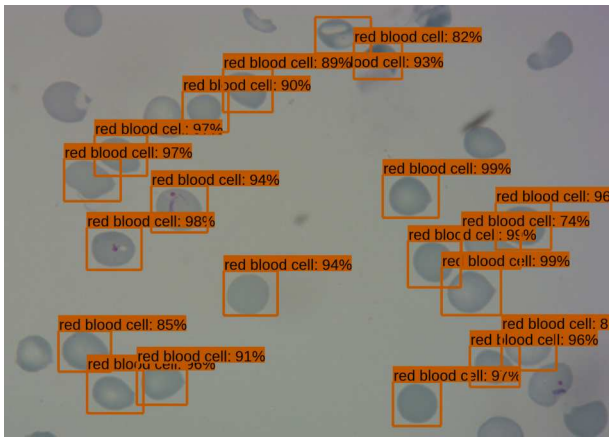

Figure 1. Sample image of Broad Institute dataset with object detection model outputs, such as bounding boxes and confidence thresholds.

Model	Accuracy	Sensitivity	Specificity	AUC	F1	MCC
ResNet50V2	0.938 ± 0.009	0.935 ± 0.012	0.940 ± 0.010	0.982 ± 0.003	0.935 ± 0.012	0.940 ± 0.014
VGG16	0.960 ± 0.003	0.956 ± 0.014	0.964 ± 0.010	0.992 ± 0.002	0.956 ± 0.014	0.964 ± 0.010
VGG19	0.959 ± 0.004	0.956 ± 0.009	0.963 ± 0.010	0.991 ± 0.001	0.955 ± 0.009	0.963 ± 0.011
InceptionV3	0.928 ± 0.001	0.925 ± 0.005	0.930 ± 0.005	0.976 ± 0.003	0.925 ± 0.005	0.930 ± 0.005
Xception	0.946 ± 0.007	0.943 ± 0.008	0.948 ± 0.010	0.979 ± 0.004	0.943 ± 0.008	0.948 ± 0.010
InceptionResNetV2	0.935 ± 0.006	0.932 ± 0.008	0.938 ± 0.007	0.980 ± 0.005	0.932 ± 0.008	0.938 ± 0.007
DenseNet121	0.956 ± 0.008	0.948 ± 0.014	0.965 ± 0.009	0.990 ± 0.003	0.948 ± 0.014	0.965 ± 0.009
MobileNetV2	0.948 ± 0.008	0.941 ± 0.012	0.955 ± 0.015	0.987 ± 0.003	0.948 ± 0.008	0.897 ± 0.016

Table 3. Transfer learning performance metrics (mean $\pm$ std). Dataset size was 10,000 images with dense nodes set to 512 with ReLU. Adam optimizer with a learning rate of 10^{-6} and batch size of 64 was used.

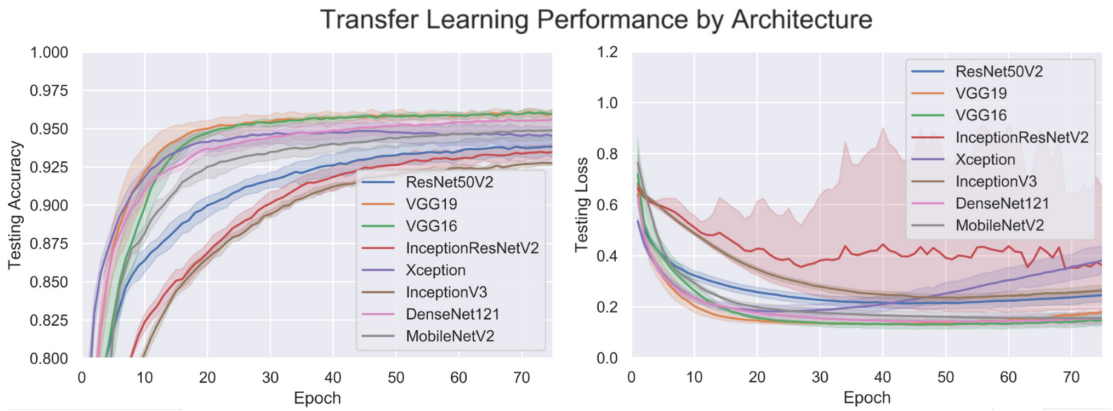

Figure 2. CNN performance with different pre-trained architectures.

Optimizing Classification Layers

Changing the number of nodes in the two dense layers after the convolution blocks does not affect the final convergence accuracy. However, increasing the number of nodes does allow the model to converge faster. Consequently, 1024 nodes were used for each dense layer during further hyperparameter tuning. A dropout rate of both 0.25 and 0.50 outperformed a dropout rate of 0.75 based on the slightly higher convergence accuracy and faster training. This suggests that a dropout rate of 0.75 may be too heavy of a regularizer. However, the dropout rate of 0.25 begins to overfit significantly more than the dropout rate of 0.50. Consequently, a dropout rate of 0.50 was used for each dense layer during further hyperparameter tuning. Lastly, the ReLU activation function appears to achieve a lower testing loss, compared to the Tanh activation function, so a ReLU activation function was used in subsequent model variants. Visualization of the effects of these hyperparameters on model training is provided in Figure 3.

Fine-Tuning Training Hyperparameters

In the first subplot in Figure 4, we see that SGD with Nesterov momentum has the fastest rise to peak accuracy, while maintaining a low testing loss even after convergence. This suggests that SGD with Nesterov momentum with a learning rate of 10^{-5} is the best optimizer to move forward with.

Image Resolution Upscaling

There is a general increase in performance of the FSRCNN model in terms of PSNR as the number of mapping convolutions (m), high resolution feature dimension (d), and low resolution feature dimension (s) increased, as shown in Table 4. The results below are derived from the most recent epoch without a dip in testing loss, as some epochs saw a temporary and drastic drop in MSE.

Settings	$m = 2$	$m = 3$	$m = 4$
$d = 48, s = 12$	30.09 (64.12)	30.07 (64.42)	30.18 (62.85)
$d = 48, s = 16$	30.30 (61.10)	30.59 (57.18)	30.72 (55.53)
$d = 56, s = 12$	30.10 (64.03)	30.25 (61.95)	30.21 (62.39)
$d = 56, s = 16$	30.42 (59.51)	30.65 (56.48)	30.79 (54.66)

Table 4. PSNR of different FSRCNN variants. MSE in parenthesis. m = number of mapping layers, d = high feature dimension space, s = low feature dimension space.

The best performing FSRCNN has a PSNR of 30.79 and a MSE of 54.66. In contrast, the traditional method of bicubic interpolation yielded a PSNR of 24.10 and a MSE of 254.67, respectively, as shown in Figure 5 with sample images. The performance values for the bicubic interpolated images are derived from the entire NIH dataset. In addition, the FSRCNN-derived images are classified more accurately than

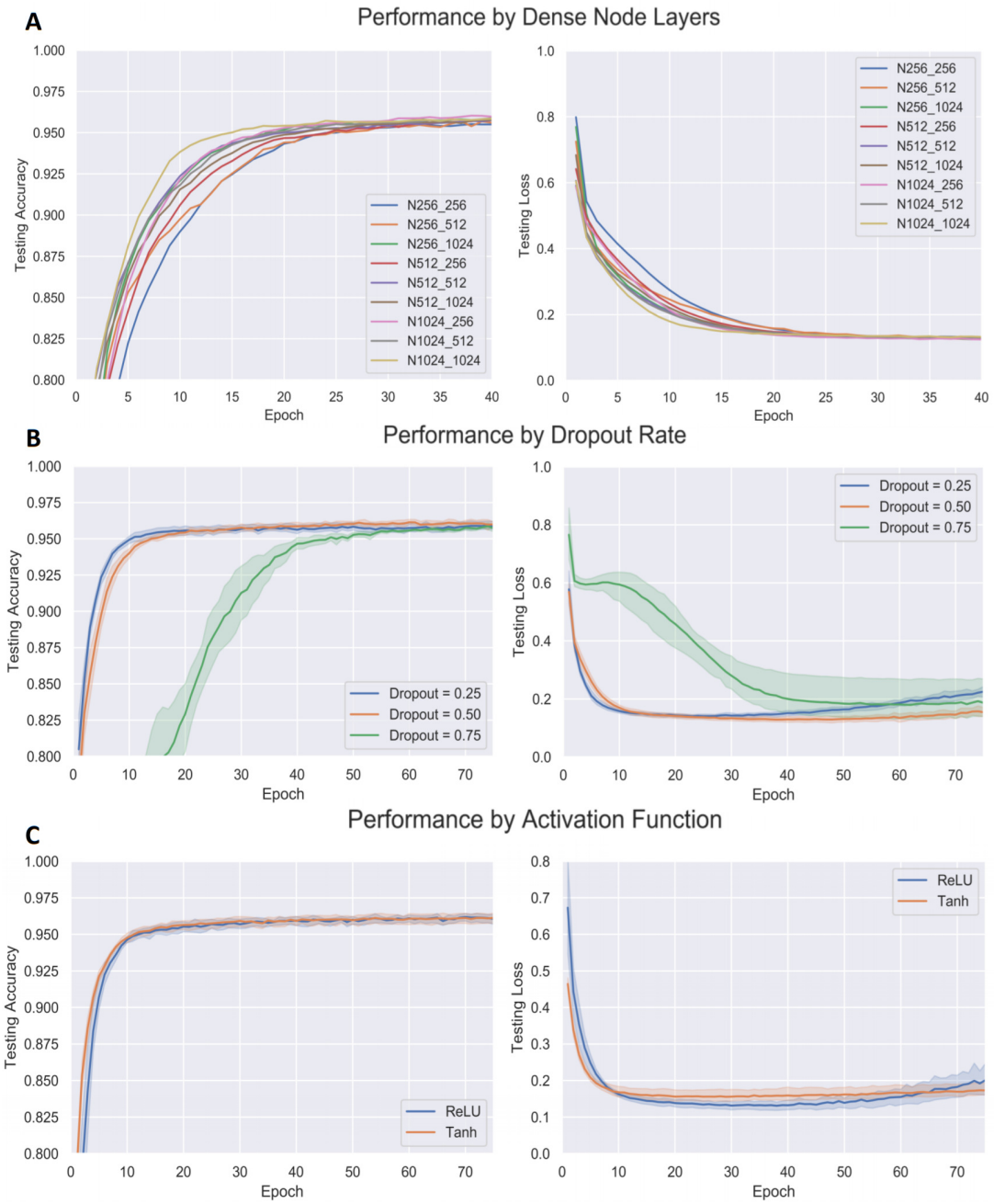

Figure 3. Performance of models with different classification layer hyperparameters. Section (**A**) displays the testing accuracy and loss with different number of nodes in each of the two dense layers. Section (**B**) displays the testing accuracy and loss with different dropout rates in the dense layers. Section (**C**) displays the testing accuracy and loss of the ReLU and Tanh activation functions in the dense layers.

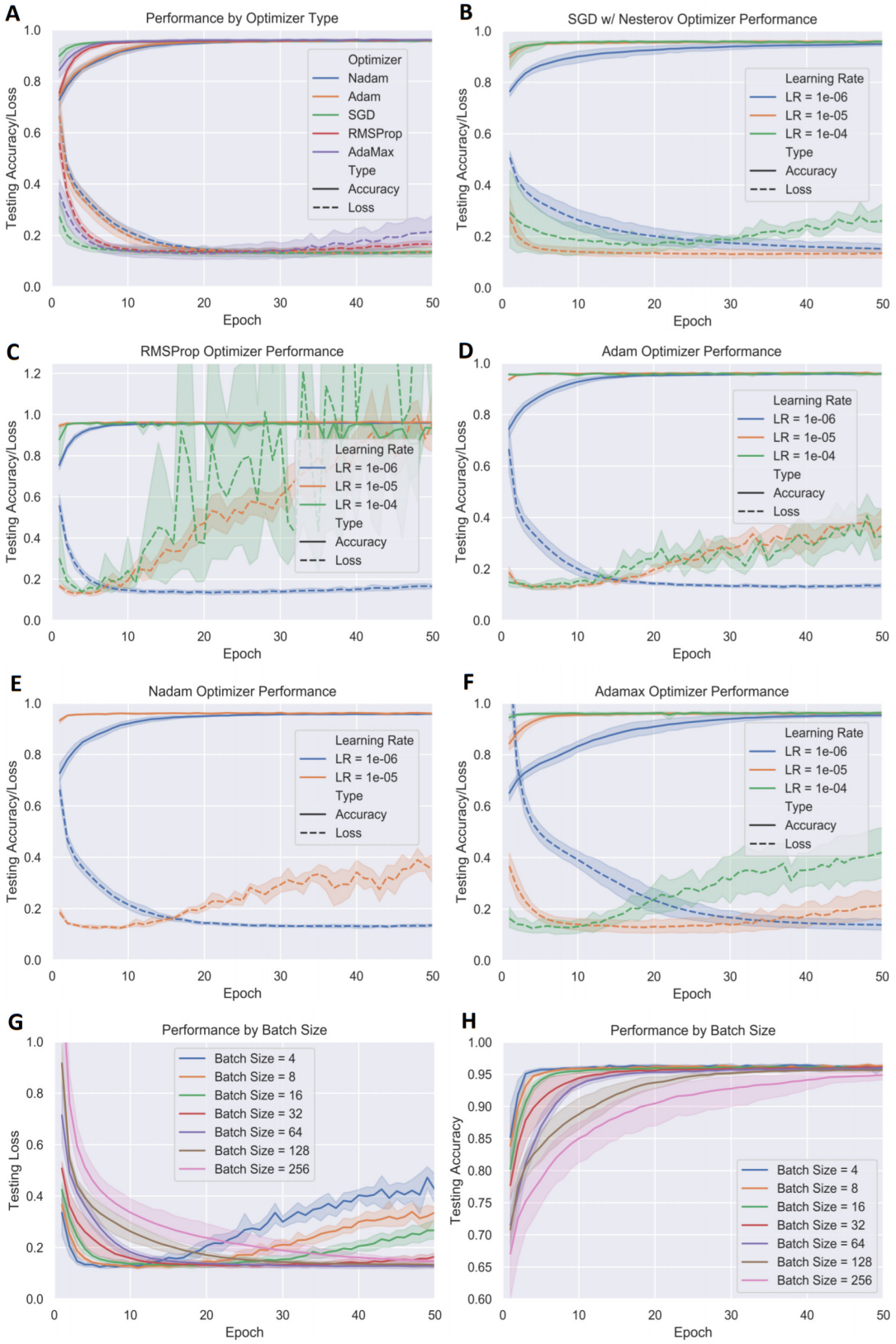

Figure 4. Performance of models with different optimizers and learning rates. Section (**A**) displays the testing accuracy and loss of the best performing learning rates of each optimizer, defined as having a fast convergence speed with minimal overfitting. Sections (**B-F**) displays the testing accuracy and loss of individual optimizers across different learning rates. Results from learning rates that resulted in a lack of improvement were omitted for clarity. Sections (**G-H**) display the testing loss and testing accuracy across different batch sizes when using a SGD w/ Nesterov optimizer with a learning rate of 10^{-5}.

the raw low resolution images or bicubic interpolated images in the finalized CNN classification model, as shown in Table 5.

Dataset	Accuracy	Sensitivity	Specificity	AUC	F1	MCC
Original High Resolution	0.9653 ± 0.0043	0.9500 ± 0.0067	0.9807 ± 0.0025	0.9940 ± 0.0010	0.9648 ± 0.0043	0.9330 ± 0.0082
FSRCNN	0.9628 ± 0.0035	0.9441 ± 0.0052	0.9815 ± 0.0027	0.9935 ± 0.0008	0.9621 ± 0.0034	0.9283 ± 0.0064
Bicubic Interpolation	0.9486 ± 0.0043	0.9093 ± 0.0106	0.9878 ± 0.0048	0.9913 ± 0.0008	0.9464 ± 0.0050	0.9022 ± 0.0078

Table 5. Classification model performance metric with different datasets (mean $\pm$ std). The original dataset contains original 128x128 images. The FSRCNN and bicubic intepolation datasets consist of downsampled 32x32 images that were rescaled upwards with their respective methods.

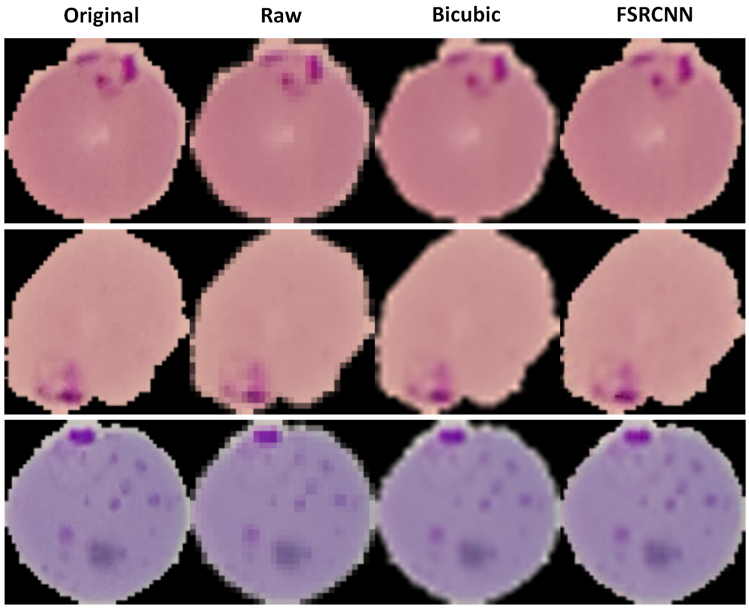

Figure 5. Sample of resolution enhanced images. Three individual *P. falciparum*-infected red blood cells from the NIH dataset. The original and upscaled images are 128x128 pixels, while the raw low-resolution image are 32x32 pixels.

Integration of CNNs on Mobile Platform

The Android app takes in an unprocessed photo of a Giemsa-stained thin blood smear, that the user manually selects on the app. Consequently, the image may either be taken directly with the phone camera or electronically acquired through other means. The SSD300 model then isolates individual images of the red blood cells and discard images of white blood cells. The image resolution of these individual images are examined so as to determine whether to upscale the image resolution via the FSRCNN model. Finally, the images are resized to 128x128 pixels and run through the VGG16 classification CNN, giving an output indicating the number of healthy and infected red blood cells, as shown in Figure 6.

DISCUSSION

Evaluation of Individual Deep Learning Components

The high average precision and relatively low average recall from the SSD300 object detection model indicates that while the detected red blood cells rarely false positives, a significant portion of red blood cells remain undetected. Because the object detection model does not distinguish between parasitized and healthy red blood cells, it is unclear whether one class of red blood cells are more likely to be go

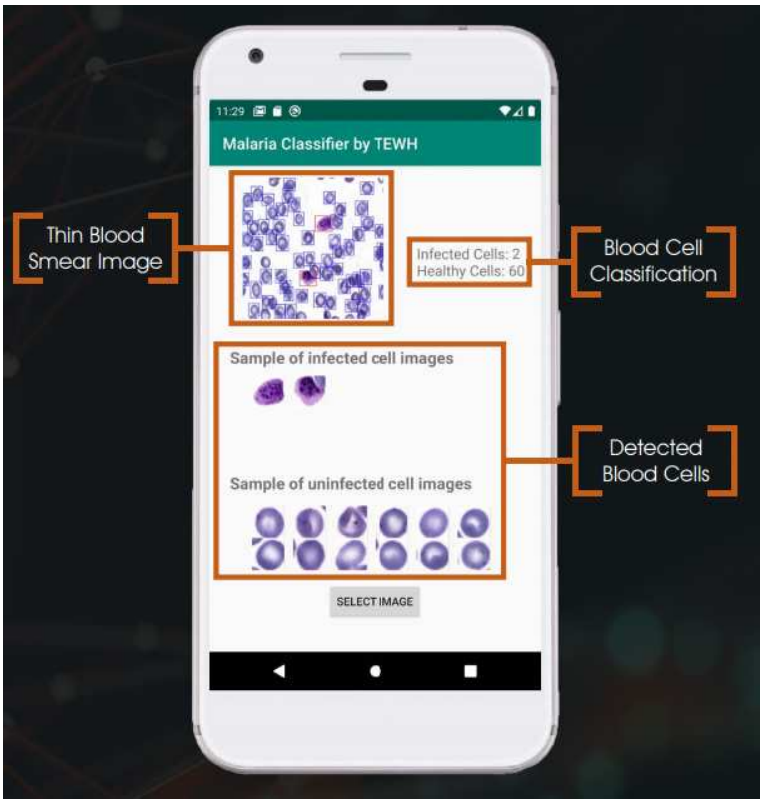

Figure 6. Example of user interface for malaria screening app. On the top left is the original thin blood smear image with the object detection bounding boxes overlaid on it. Individual images of red blood cells, as well as cell counts, are provided as well.

undetected by the SSD300 model. However, it would be ideal that both parasitized and uninfected red blood cells are equally likely to be detected by the object detection model, because the severity of a malaria infection is often measured in percent parasitemia, or the percentage of infected red blood cells.

In the FSRCNN image upscaler, we see while the resolution enhancement process generates significant improvements in the CNN classification model performance, compared to the traditional scaling method bicubic interpolation. This shows that even for simplistic structures such as red blood cells, low-resolution images will cause the classification model to perform significantly more poorly. This is a critical consideration to keep in mind, as image resolution may be limited during the image acquisition process if the camera has poor resolution. Additionally, we see that increasing the number of mapping layers, the high resolution feature dimension, and low resolution feature dimension, all tend to promote an increase in the effectiveness of resolution upscaling. However, it is worth noting that the central purpose of the FSRCNN model is to demonstrate whether improved resolution upscaling methods can positively impact subsequent classification. Recent developments suggest that the use of novel GANs - such as the SRGAN - yield better PSNR results, and may be a better models to implement during further development (Christian Ledig, 2017).

Meanwhile, our classification CNN model has an accuracy of about 96.53% and an AUC of 0.994, which is lower than the accuracies of other groups who have also trained their model on the NIH dataset. However, it is worth noting that the highest performance reported by (Sivaramakrishnan Rajaraman, 2019) were due to the use of ensemble networks, which may not be feasible for mobile phone use due to its heavier computational burden. Meanwhile, the highest performance reported by (Aimon Rahman, 2019) was from a model trained on a modified NIH dataset, in which the group reports that incorrectly labeled images were removed from the dataset prior to training. Top performing non-ensemble models reported by (Zhaohui Liang, 2017) and (Sivaramakrishnan Rajaraman, 2018) report classification accuracies of about 97.4% and 98.6%, respectively. However, neither groups tested their final models on a separate independent dataset to examine the generalizability of their models. The performance of our NIH dataset-trained classification model significantly dropped when tested on the Broad Institute dataset, with AUC of 0.945 ± 0.025, compared to an AUC of 0.994 ± 0.001 with the cross-validated NIH dataset. This suggests that the current classification model does not generalize well towards due to the three following differences between the NIH and Broad Institute datasets: (1) unsegmented vs segmented images, (2) *P. falciparum* vs *P. vivax* parasites, and (3) overlapping vs non-overlapping cells in individual images.

Eliminating the Need for Internet Access and Manual Segmentation in the Mobile App

We present a proof-of-concept with our streamlined, mobile phone-powered screening platform. A flexible Android app framework has been developed, in which any of the model components can be easily removed and replaced with an new and higher-performing model. Additionally, the code outside of the .tflite files within the Android app is basic and brief, performing basic tasks such as transferring the outputs of the resolution upscaling model to the classification model for diagnostic results. While other groups such as (Sivaramakrishnan Rajaraman, 2018) have reported similarly designed mobile phone apps, the apps transmit images to a cloud-based model for classification. This poses an additional barrier in areas with low or non-existent mobile phone internet connectivity. To our knowledge, our phone app is the only malaria screening app that is currently reported to run entirely on the mobile phone without the need for internet access. In addition, our mobile phone app requires only a thin blood smear image, rather than already segmented images of each individual red blood cell.

Immediate Barriers to Deployment

The two major barriers towards employing the phone-based deep learning models are: (1) the lack of a comprehensive malaria blood smear dataset and (2) the generalizability of the models themselves.

Lack of Comprehensive Dataset The NIH dataset contains images of individual *P. falciparum*-infected red blood cells that are already segmented. Meanwhile, the Broad Institute dataset contains images of *P. vivax*-infected red blood cells with bounding boxes but no segmented images. Consequently, this results

in a dilemma for realistic application in developing countries. In order to effectively utilize a classification CNN trained on segmented images, we must develop a corresponding cell segmentation model. However, the lack of a dataset with both segmented and unsegmented images makes it impossible to develop such a model. This is problematic for our current models, in which the SSD object detection model was trained for object detection rather than image segmentation, while the classification model was trained on segmented images. Alternatively, a classification CNN could be trained on unsegmented images and only bound images of individual red blood cells, as seen in the Broad Institute dataset. However, the Broad Institute dataset contains *P. vivax* parasites, rather than the predominant and deadlier *P. falciparum* parasites found in African regions. Consequently, an important immediate objective is to acquire a comprehensive dataset that alleviates these issues.

Generalizability of Deep Learning Models Although *P. falciparum* accounts for the majority of malaria infections in African regions, *P. vivax* is indeed the second most common parasite. In a low-resource setting, it is difficult if not impossible to discern which specific parasite is present in a thin-blood smear outside of manual observation of the thin blood smears. Consequently, an important improvement over current advances would be developing a generalizable deep learning model that is able to indiscriminately detect malaria-infected red blood cells, regardless of the specific parasite present. It seems that no group has attempted this yet. Lastly, as seen in the Broad Institute dataset, there is often significant overlap between individual red blood cells, which may interfere with the accuracy of our current classification model, which was trained on non-overlapping individual red blood cells.

CONCLUSIONS

While many groups have attempted to use machine learning algorithms to automate the detection and classification of malaria-infected red blood cells, there has not been significant effort towards object detection and image resolution upscaling in the context of the malaria screening process.

By introducing a proof-of-concept, with a preliminary SSD300 object detection model and FSRCNN resolution upscaling model in tandem with a single-cell classification model, we show that a streamlined and sequential approach towards automating the diagnosis of malaria from input of the blood smear to output of the number of infected and healthy red blood cells may be possible as the individual models are further developed.

With the rapid advancements made every year in deep learning technology, faster and more accurate models developed in the near future can easily be switched with the models used our phone app due to the modularity of our code. This allows us to move closer towards real implementation in developing countries without the need for trained technicians or internet-based computing resources.

ACKNOWLEDGEMENTS

To be included if manuscript is accepted: We would like to thank our anonymous reviews for their helpful feedback in the review process of this manuscript.

SUPPLEMENTAL INFORMATION

Competing Interests

The authors declare that there are no competing interests.

Author Contributions

Oliver S. Zhao conceived and designed the experiments, performed the classification based experiments, prepared figures and tables, and wrote the manuscript.

Nikhil Kolluri, Annie Anand, and Nicholas Chu implemented the object detection model experiments, developed the TensorFlow Lite platform for mobile implementation of all models, and managed the Google Cloud Computing platform that was used to run experiments.

Ravali Bhavaraju and Sandhya Tiku helped fine-tune classification models and preprocess images, in addition to helping prepare figures based on results from other experiments.

Aditya Ojha and Ryan Chen implemented the image resolution upscaling method and ran experiments characterizing its performance under different conditions.

Dat Nguyen, Adriane Morales, Deepti Valliappan, Juhi Patel, and Kevin Nguyen aided in the fine-tuning of the classification model hyperparameters.

Data Availability

The following publicly available datasets that were used can be found at the following sites:

NIH NLM Dataset: https://lhncbc.nlm.nih.gov/publication/pub9932

Broad Institute Dataset: https://data.broadinstitute.org/bbbc/BBBC041/

Supplemental information containing results derived from the experiments outlined in this manuscript are publicly available at: https://github.com/oliver29063/MalariaDiagnosis

Funding

The research outlined was funded by donations provided to Texas Engineering World Health (TEWH), a student-chapter of the parent organization Engineering World Health, based at The University of Texas at Austin. Individual donors and other TEWH members that are not listed on the authorship list had no role in any part of the research or writing of the manuscript.

REFERENCES

Aimon Rahman, Hasib Zunair, M. S. R. J. Q. Y. S. B. M. A. A. N. B. A. M. M. (2019). Improving malaria parasite detection from red blood cell using deep convolutional neural networks. *ArXiv*.

CDC (Accessed 05-07-2020). Malaria diagnosis and treatment in the united states. Technical report, Centers for Disease Control and Prevention.

Chao Dong, Chen Change Loy, K. H. X. T. (2014). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307.

Chao Dong, Chen Change Loy, X. T. (2016). Accelerating the super-resolution convolutional neural network. *ArXiv*.

Christian Ledig, Lucas Theis, F. H. J. C. A. C. A. A. A. A. A. T. J. T. Z. W. W. S. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE CVPR*, pages 105–114.

Dev Kumar Das, Madhumala Ghosh, M. P. A. K. M. C. C. (2013). Machine learning approach for automated screening of malaria parasite using light microscopic images. *Micron*, 45:97–106.

Dinggang Shen, Guorong Wu, H.-I. S. (2017). Deep learning in medical image analysis. *Annu Rev Biomed Eng*, 19:221–248.

Duangporn Polpanich, Pramuan Tangboriboonrat, A. E. R. U. (2007). Detection of malaria infection via latex agglutination assay. *Anal Chem*, 79:4690–4695.

E. Charpentier, E. Benichou, A. P. P. C. J. F. A. V. H. G. E. G. A.-S. S. C. A. S. M. S. C. A. B. X. I. (2020). Performance evaluation of different strategies based on microscopy techniques, rapid diagonstic test and molecular loo-mediated isothermal amplification assay for the diagnosis of imported malaria. *Clin Microbiol Infect*, 1:115–121.

F. Ellis McKenzie, David L. Smith, W. P. O. E. M. R. (2008). Strain theory of malaria: The first 50 years. *Adv Parasitol*, 66:1–46.

Geert Litjens, Thijs Kooi, B. E. B. A. A. A. S. F. C. M. G. J. A. d. L. B. v. G. C. I. (2017). A survey on deep learning in medical image analysis. *Med Image Anal*, 42:60–88.

Gopalakrishna Pillai Gopakumar, Murali Swetha, G. S. S. G. R. K. S. S. (2017). Convolutional neural network-based malaria diagnosis from focus stack of blood smear images acquired using custom-built slide scanner. *Journal of Biophotonics*, 11: Epub.

Haris M Khan, Fatima Shujatullah, M. S. A. R. R. M. (2010). Evaluation of diagnos malaria stix test (antigen detection assay) for diagnosis of malaria. *J Commun Dis*, 42:153–156.

Kaiming He, Xiangyu Zhang, S. R. J. S. (2016). Deep residual learning for image recognition. *2016 IEEE CVPR*.

Kan, A. (2017). Machine learning applications in cell image analysis. *Immunol Cell Biol*, 95:525–530.

Karen Simonyan, A. Z. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR 2015*.

Katherine M. Bowers, David Bell, P. L. C. J. B. S. I. S. Y. J. L. H. W. (2009). Inter-rater reliability of malaria parasite counts and comparison of methods. *Malar J*, 8.

Kristofer E. Delas Peñas, Pilarita T. Rivera, P. C. N. J. (2017). Malaria parasite detection and species identification on thin blood smears using a convolutional neural network. *IEEE CHASE 2017 Proceedings*.

Kusworo Adi, Sri Pujiyanto, R. G. A. P. A. B. P. (2016). Identifying the developmental phase of plasmodium falciparum in malaria-infected red blood cells using adaptive color segmentation and back propagation neural network. *IJAER*, 11:8754–8759.

Linda M. Murungi, Rinter K. Kimathi, J. T. G. K. F. H. A. O. (2019). Serological profiling for malaria surveillance using a standard elisa protocol. *Methods Mol Biol*, pages 83–90.

Mounkaila Abdou Billo, Mahamadou Diakité, A. D. M. D. B. P. S. I. D. E. S. J. J. C. R. D. J. K. O. K. D. (2013). Inter-observer agreement according to malaria parasite density. *Malar J*, 12.

Nguyen Van Hong, Peter van den Eede, C. V. O. I. V. A. R.-U. P. V. T. N. D. T. N. M. H. L. X. H. U. D. and Erhart, A. (2013). A modified semi-nested multiplex malaria (snm-pcr) for the identification of the five human plasmodium species occurring in southeast asia. *Am J Trop Med Hygn*, 89:721–723.

Nicholas E. Ross, Charles J. Pritchard, D. M. R. A. G. D. (2006). Automated image processing method for the diagnosis and classification of malaria on thin blood smears. *Medical and Biological Engineering and Computing*, 44:427–436.

Rasheed O. Makanjuola, A. W. T.-R. (2020). Improving accuracy of malaria diagnosis in underserved rural and remote endemic areas of sub-saharan africa: A call to develop multiplexing rapid diagnostic tests. *Scientifica (Cairo)*, page ePub.

Ricci, F. (2012). Social implications of malaria and their relationships with poverty. *Mediterr J Hematol Infect Dis*, 4:ePub.

Shidan Wang, Donghan M. Yang, R. R. X. Z. X. Z. G. X. (2019). Pathology image analysis using segmentation deep learning algorithms. *Am J Pathol*, 9:1686–1698.

Sivaramakrishnan Rajaraman, Sameer K. Antani, M. P. K. S. M. A. H. R. J. M. S. J. G. R. T. (2018). Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. *PeerJ*, Epub.

Sivaramakrishnan Rajaraman, Stefan Jaeger, S. K. A. (2019). Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *PeerJ*, Epub.

Syied Anwar, Muhammad Majid, A. Q. M. A. M. A. K. K. (2018). Medical image analysis using convolutional neural networks: A review. *J Med Syst*, 42:226.

WHO (2019). World malaria report 2019. Technical report, World Health Organization.

Yuhang Dong, Zhuocheng Jiang, H. S. W. D. P. L. A. W. V. V. R. W. H. B. A. W. B. (2017). Evaluations of deep convolutional neural networks for automatic identification of malaria infected cells. *IEEE EMBS 2017 Proceedings*, pages 101–104.

Zhaohui Liang, Andrew Powell, I. E. M. P. K. S. K. P. P. G. M. A. H. A. S. R. J. M. J. X. H. S. J. G. T. (2017). Cnn-based image analysis for malaria diagnosis. *IEEE BIBM 2016 Proceedings*, pages 493–496.