Applying Convolutional Neural Networks to Identify Parasitized Malaria Cells

Rohan Bhansali rohanmbhansali@gmail.com
Rahul Kumar rakumar2003@gmail.com

1. About the authors

Rohan Bhansali is a junior in high school that is conducting deep learning research at the Academy of Science in Loudoun County, Virginia. He is passionate about working with deep learning algorithms to solve biomedical problems to help benefit the less fortunate citizens of the world.

Rahul Kumar is a junior in high school working on machine learning research at the Academy of Science in Loudoun County, Virginia. He is ardent about using technology to help solve problems in innovative and efficient ways to help people around the world.

2. Abstract

We develop an algorithm that can detect malaria from images of segmented cells from the thin blood smear slide images with 96% accuracy. Our algorithm, SimpNet-7, is a 7-layer convolutional neural network trained on the NIH Malaria dataset, containing 27,588 images of parasitized and uninfected cells. We find that SimpNet-7 achieves an F1 score of 0.955, a precision of 0.946, and a recall of 0.974. We then propose the application of this algorithm in hospitals in areas where malaria is prominent but medical resources are sparse, such as African countries. Our source code and pretrained weights are available here.

3. Introduction

Malaria is a mosquito-borne infectious disease that is caused by single-celled, parasitic microorganisms of the *Plasmodium* genus and is typically spread via the bite of infected female *Anopheles* mosquitoes, as the parasites from the saliva of the mosquito are transferred to the blood of the affected human. Subsequently, the parasites transfer to the liver where they are able to efficiently develop and reproduce. The disease manifests itself as fever, tiredness, vomiting, and headaches; in severe cases, it can also result in yellow skin, seizures, coma, and even death. These symptoms typically present themselves ten to fifteen days after infection, but recurrence can occur months later. Although medicinal research has made strides in malaria prevention, detection, and alleviation, malaria remains the leading cause of death in the world and at this time, there exists no effective vaccination. Most cases occur in tropical and subtropical regions, particularly Sub-Saharan Africa, Asia, and Latin America. Overall, malaria was estimated to affect 228 million people and cause an estimated 405,000 deaths in 2018, with 94% of these fatalities occurring in Africa. Consequently, the disease is estimated to cost Africa over $12 billion due to healthcare costs, decreased workforce, and negative effects on tourism.

The most widely used methods for malaria diagnosis fall into two categories: direct and indirect. Direct methods of detection rely upon confirming a diagnosis based upon the discovery of parasitic bodies or parts of parasitic bodies. Indirect methods involve the detection of antibodies known to be relevant agents in the diagnosis of malaria. The advantages and disadvantages of the primary direct and indirect malaria diagnostic test methods were recorded by Talapko and colleagues and are re-expressed in *Table 1* and *Table 2*.

Table 1 (Talapko et al., 2019)

Direct Methods	Advantages	Disadvantages
Microscopic Analysis	Quick and inexpensive	Requires high-level experience and training
Rapid Diagnostic Tests	Quick and simple to perform	Low sensitivity and expensive
Molecular Tests	Highly sensitive and accurate	Expensive and time-consuming in a large fraction of cases

Table 2 (Talapko et al., 2019)

Indirect Methods	Advantages	Disadvantages
Indirect Immunofluorescence	High specificity and sensitivity	Time-consuming to perform and subjective evaluation of results
ELISA	Correct determination of type, highly specific	Time-consuming to perform and expensive

As seen in both *Table 1* and *Table 2*, the current methods of Malaria diagnosis all feature a combination of issues in price, accessibility, evaluation of results, or low levels of sensitivity.

According to the CDC, microscopic analysis of potentially infected cells remains the gold standard in malaria diagnosis. However, they note that the quality and accuracy of the test is highly dependent on the experience of the laboratorian conducting the diagnostic (CDC, 2019). As a result, Malaria detection through microscopic analysis is inaccessible to millions around the world since there are

limited numbers of experienced laboratorians in many African and Asian countries. For these reasons, despite being one of the cheapest and quickest methods of Malarial diagnosis, microscopic analysis is still unattainable for the majority of high-risk individuals across the world. For this reason, the purpose of this study is to develop a robust computational algorithm with the ability to classify malaria afflicted cells with a high degree of accuracy.

4. Data

1. *Dataset Acquisition*

The dataset used in this study was provided by the National Institute of Health (NIH) and contained 27,558 images with equal instances of parasitized and non-parasitized red blood cells. The cells were stained with Giemsa, mimicking the procedure that would be undertaken when using microscopic analysis to diagnose malaria. The dataset is described here.

2. *Image Preprocessing*

The first step was to process the images and split them into train and test subsets, thereby making the dataset ready for usage. The images were processed via the nearest-neighbor interpolation technique, in which a group of pixels automatically assume the magnitude of the pixel closest to it. It is the simplest method of multivariate interpolation, which also makes it the most computationally efficient. The images were downscaled to a standard size of 100x100 pixels, which was necessary as the images were previously of variable size and scale. Applying nearest-neighbors interpolation minimized image distortion and retained information better than less sophisticated methods such as resizing or cropping. We then placed 80% of the images in a training dataset, while the remaining 20% were used for testing.

5. Model architectures

We utilized Orange3, an open-source machine learning and data mining application, with our newly processed images. This served two purposes; not only did it allow us to determine the fitness of our images for machine learning classification, but it also allowed us to easily test and compare different algorithm architectures against one another. We utilized four different types of models: neural network, logistic regression, k-nearest neighbors, and random forest. As our metrics, we used each model's confusion matrix, AUC score, F1 score, precision and recall, sensitivity and specificity, and accuracy for comparison.

Table 3: Model comparison statistics

Model	AUC	F1	Precision	Recall	Sensitivity	Specificity	Accuracy
kNN	0.938	0.866	0.874	0.867	0.928	0.821	86.679%
Logistic Regression	0.983	0.941	0.941	0.941	0.945	0.937	94.103%
Neural Network	0.985	0.944	0.944	0.944	0.945	0.944	94.419%
Random Forest	0.958	0.895	0.896	0.896	0.878	0.915	89.552%

Table 4: kNN confusion matrix

		Predicted		
		Parasitized	Uninfected	∑
Actual	Parasitized	10957	2822	13779
	Uninfected	849	12930	13779
	∑	11806	15752	27558

Table 5: Logistic regression confusion matrix

		Predicted		
		Parasitized	Uninfected	∑
Actual	Parasitized	12910	869	13779
	Uninfected	756	13023	13779
	∑	13666	13892	27558

Table 7: Neural network confusion matrix

		Predicted		
		Parasitized	Uninfected	∑
Actual	Parasitized	13001	778	13779
	Uninfected	760	13019	13779
	∑	13761	13797	27588

Table 8: Random Forest confusion matrix

		Predicted		
		Parasitized	Uninfected	∑
Actual	Parasitized	12657	1122	13779
	Uninfected	1757	12022	13779
	∑	14414	13144	27588

As the statistics demonstrate, neural networks proved to be the superior model. Convolutional neural networks, a specific type of neural network, are especially strong at image classification due to their ability to maintain information quality while concurrently reducing parameters, boosting efficiency. As a result, we decided to employ this type of algorithm in classifying the red blood cells.

6. SimpNet-7

The malaria detection task is a binary classification problem, where the input is an image of a cell X and the output is a binary label $y \in \{0, 1\}$ indicating the absence or presence of malaria, respectively.

To accomplish this task, a seven-layer CNN was utilized that we dubbed SimpNet-7. This model contained two convolution layers, two pooling layers, a flattening layer, and two fully connected dense layers. The dense layers alleviate the vanishing gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. The network structure can be seen in *SimpNet-7 Layer Map.* Within our model, we utilized Adaptive Moment Estimation (Adam), an adaptive learning rate optimization algorithm. Our activation functions for the intermediate layers was the rectifier function, with the final layer using a sigmoid nonlinearity. A batch size of ten was used over 25 epochs, after each of which the

model output accuracy metrics and loss, given by the binary cross-entropy loss function: $L(X,y) = -w+\cdot ylogp(Y=1|X) - w \cdot(1-y)logp(Y=0|X)$, where p(Y = i|X) is the probability that the network assigns to the label i, w+ = |N|/(|P|+|N|), and w− = |P |/(|P |+|N |) with |P | and |N | the number of parasitized and uninfected cells in the training set respectively.

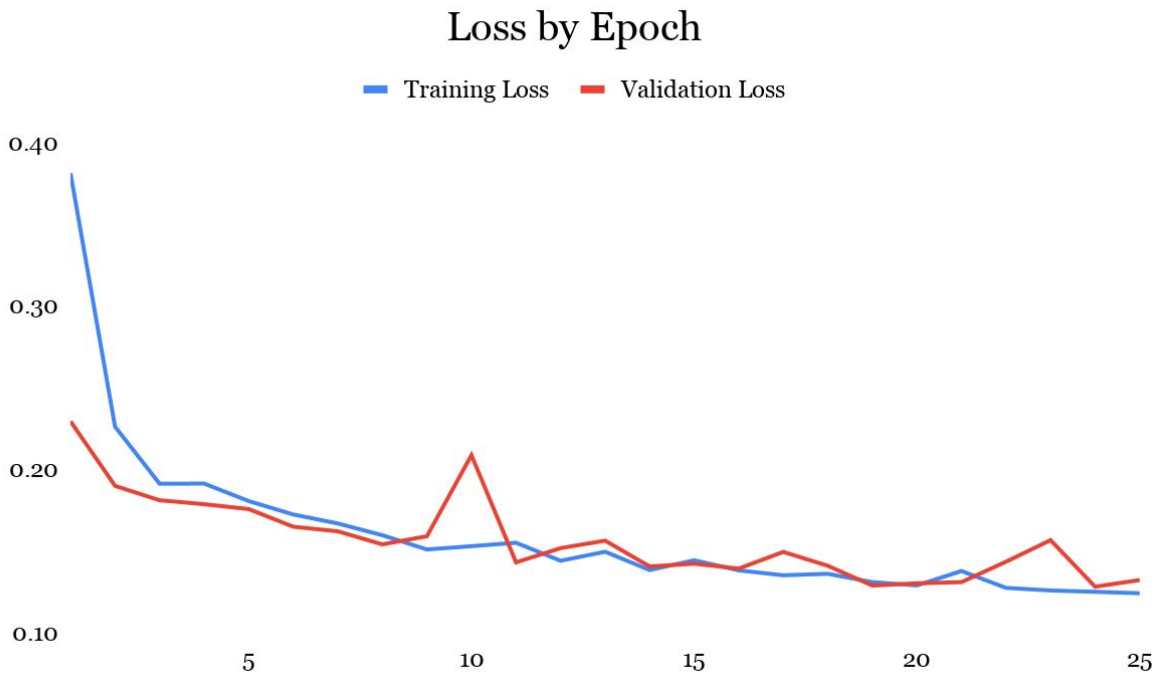
In order to prevent the networks from overfitting, early stopping was performed by saving the network after every epoch and choosing the saved network with the lowest loss on the tuning set. Overall, 2,177,185 parameters were trained and optimized for this task. The architecture of the network can be seen in *Table 9.*

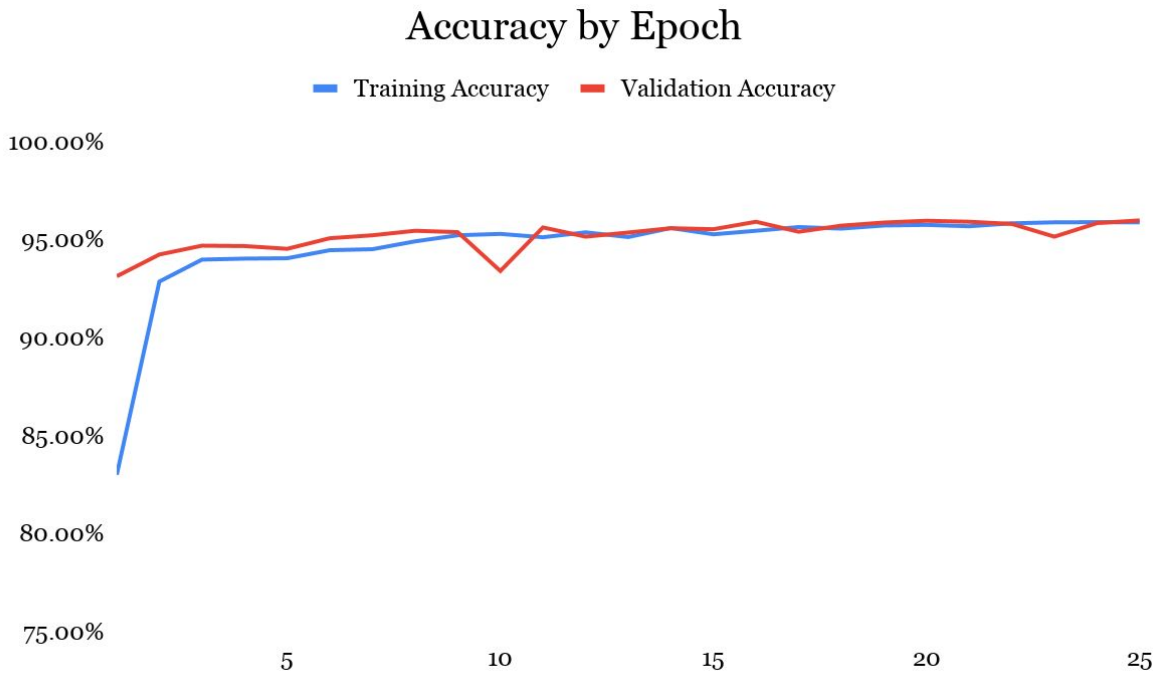
Table 9: SimpNet-7 Architecture

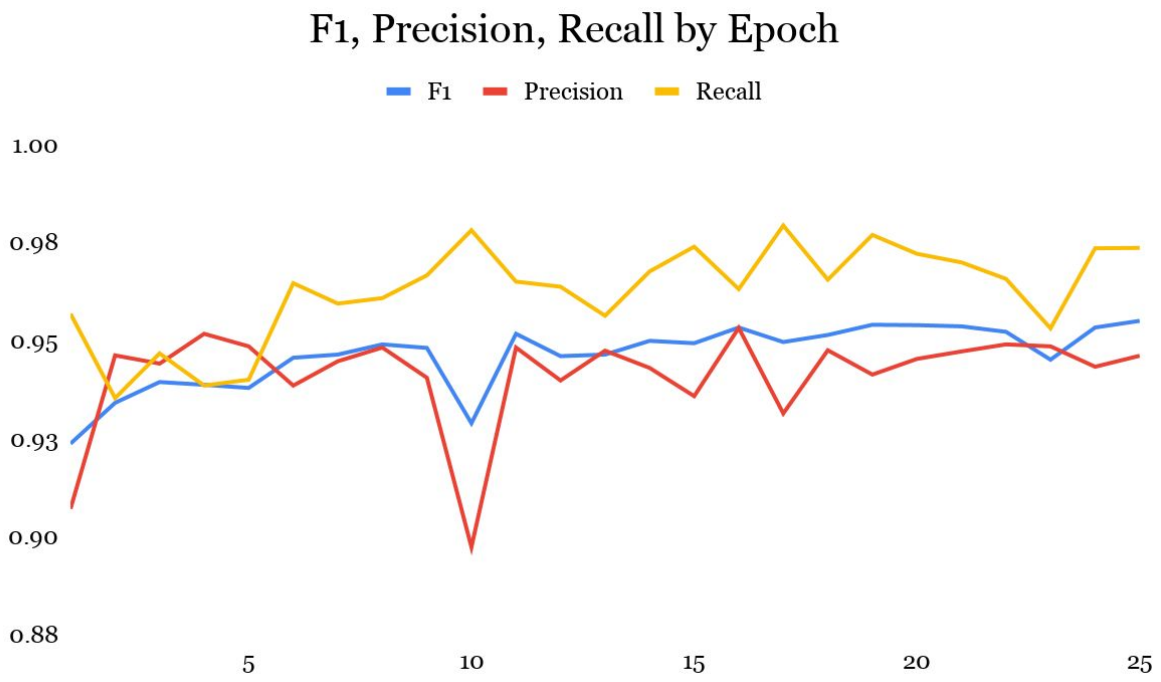
Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 98, 98, 32)	896
Max_pooling2d_1 (MaxPooling2)	(None, 49, 49, 32)	0
conv2d_2 (Conv2D)	(None, 47, 47, 32)	9248
max_pooling2d_2 (MaxPooling2)	(None, 23, 23, 32)	0
flatten_1 (Flatten)	(None, 16928)	0
Dense_1 (Dense)	(None, 128)	2166912
Dense_2 (dense)	(None, 1)	129

7. Results

The progression of SimpNet-7 training can be seen below. We present loss, accuracy, F1 score, precision, and recall by each epoch.







SimpNet-7 achieved relatively high performance in all metrics after training was complete, as can be seen in *Table 10*.

Table 10

Accuracy	F1	Precision	Recall
0.960	0.955	0.946	0.974

8. Conclusion

Malaria is a disease that afflicts hundreds of thousands of individuals around the world. A primary form of diagnosis for the disease utilizes microscopic analysis, which currently requires trained laboratories for analysis and operation that are not abundantly available around the world. To help alleviate this, we developed a CNN

architecture that was trained on a publicly available dataset containing approximately 27,000 images of red blood cells in order to classify malaria. We achieved an F1 score of 0.955 with our deep learning model, making it highly appropriate for medical implementation around the world. SimpNet-7 has potential to provide groundbreaking diagnostic capabilities to the areas where malaria is most prevalent around the world. We hope that our technology can be used in remote areas to help save the lives of thousands of people that are at risk of contracting the disease.

9. References

Analysis of research and development priorities for ... (n.d.). Retrieved from https://www.who.int/research-observatory/analyses/malaria_rd_priorities_working_paper.pdf?ua=1

CDC - Parasites - Malaria. (2020, April 1). Retrieved from https://www.cdc.gov/parasites/malaria/index.html

Improving Malaria Parasite Detection from Red Blood Cell ... (n.d.). Retrieved from https://www.researchgate.net/publication/334669002_Improving_Malaria_Parasite_Detection_from_Red_Blood_Cell_using_Deep_Convolutional_Neural_Networks

Malaria Datasets | National Library of Medicine. (n.d.). Retrieved from https://lhncbc.nlm.nih.gov/publication/pub9932

Narayanan, B. N., Ali, R., & Hardie, R. C. (2019, September 6). Performance analysis of machine learning and deep learning architectures for malaria detection on cell images. Retrieved from https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11139/2524681/Performance-analysis-of-machine-learning-and-deep-learning-architectures-for/10.1117/12.2524681.short?SSO=1

Rahman, A. (2019). Improving Malaria Parasite Detection from Red Blood Cell using Deep Convolutional Neural Networks. DeepAI.

Shi, G. (2019, January 30). Detecting malaria using deep learning. Retrieved from https://towardsdatascience.com/detecting-malaria-using-deep-learning-fd4fdcee1f5a

Talapko, J., Škrlec, I., Alebić, T., Jukić, M., & Včev, A. (2019, June 21). Malaria: The Past and the Present. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6617065/