JOURNAL OF
EMERGING INVESTIGATORS

Article

Machine learning for the diagnosis of malaria: a pilot study of transfer learning techniques

Jacob E. Stolker[1,2] and Rida Assaf[1,3]

[1]Horizon Academic Research Program, Jersey City, NJ
[2]Computer Science, Grainger College of Engineering, University of Illinois, Urbana-Champaign, IL
[3]Department of Computer Science, University of Chicago, Chicago, IL

SUMMARY

Despite being treatable and preventable, malaria continues to affect millions of people worldwide. A major hindrance to the disease's eradication is difficulty with testing, and attempts have already been made to shift towards more automated testing methods rather than older testing technologies such as light microscopy and rapid diagnostic tests. Machine learning models have shown promise in diagnosing malaria, but prior machine learning algorithms have been contingent upon the availability of clean and relevant data. Furthermore, there are multiple species of malaria that may require different treatments, often resulting in fewer comparative images of individual malarial subtypes on blood slides of a given patient population. In this pilot study, we tested a machine learning technique called transfer learning to reduce the error caused by lesser amounts of available data. By training the algorithm on a known (but not necessarily related) dataset and then re-training on malaria datasets, we were able to significantly increase diagnostic accuracy. Specifically, using transfer learning to prepare the algorithm on a completely unrelated database before applying to malaria slide interpretation, our diagnostic accuracy improved from 65.5% to 80.7% in a small dataset of malaria blood slides, and from 95.5% to 98.0% in a much larger malarial dataset. Moving forward, these types of experiments will help develop a smartphone application to increase the opportunity to diagnose malaria in places where resources are too limited to permit standard diagnostic techniques.

INTRODUCTION

According to the World Health Organization, around 228 million people were affected by malaria in 2018 (1). Malaria is caused by parasitic Plasmodium species that are transmitted to humans by female mosquitoes during blood meals. There are five species of Plasmodium that affect humans, each of which has a different life cycle and has a different microscopic appearance on a blood slide (2). Despite medical advances allowing this disease to be more rapidly treated, lack of quality testing has contributed to the continued devastation of mostly equatorial countries, especially those in sub-Saharan Africa (3).

One major issue preventing the eradication of malaria is the poor infrastructure of remote regions and sub-par training of personnel at smaller testing centers (4). In addition, many centers lack personnel skilled at diagnosing malaria and/or adequate medical equipment to perform the tests. The reason for this shortage is the high degree of training and cost of equipment required to perform the tests. To test for malaria through light microscopy, a skilled medical worker must prepare and observe a blood slide through a high-power microscope. Then, the worker must carefully observe the slide and make the best call on whether there is a presence of malaria, which can be costly and inaccurate. Rapid Diagnostic Tests (RDTs) function differently, requiring just a prick of a patient's blood to test for malaria antibodies (5). These tests have a different set of issues, as RDTs are more expensive while only detecting certain species of Plasmodium, and often provide less accurate information regarding the diagnosis as a result (6).

Given these limitations, one modern solution that has been applied before involves the implementation of a process called image analysis, a slightly more primitive process compared to machine learning for the diagnosis of malaria. The potential benefits of image analysis are less cost and increased accuracy. With the diagnosis of malaria blood slides, a process called segmentation must be applied before image analysis can be performed. Through image analysis, some of the diagnostic problems can be solved by providing more accurate and precise results through automated analysis of blood slide images. The traditional method of diagnosing malaria based on image analysis requires two primary steps: first, segmentation of the blood slide into viewable sections, and second, the direct detection of Plasmodium particles in the sections of the microscope slide. Segmentation is a process where the image is enhanced and analyzed through hue, saturation, and intensity to reveal infected cells. This is done by adding some sort of buffer solution to samples to reveal certain characteristics to analyze. However, different factors, such as inconsistent use of variable buffer solutions on blood slides, can alter these values, thus making the interpretation of the slides inaccurate. Since segmentation and image analysis involve finding patterns manually and or directly creating an algorithm to detect patterns to classify images, the predictions are not as accurate as the machine learning algorithms that have largely replaced this labor-intensive, time-consuming, and costly process (7).

We used a type of machine learning called categorical classification, in which a computer model is trained on pre-

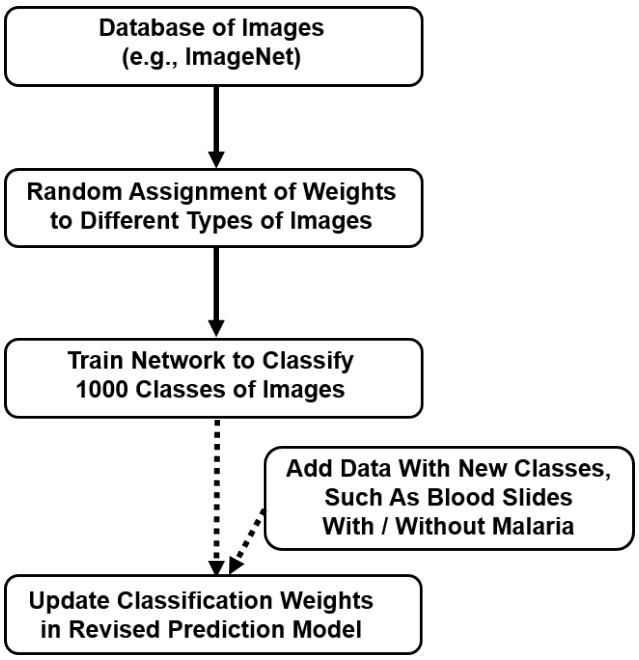

Figure 1: Diagram of the transfer learning process.
In this study, we used an InceptionV3 neural network pretrained on ImageNet for the transfer learning tests.

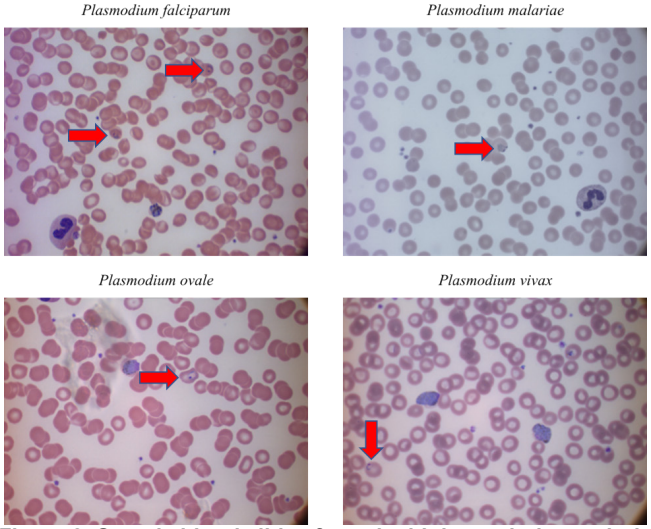

Figure 2: Sample blood slides from the high-resolution malaria dataset.
Example blood slides of *P. falciparum*, *P. malariae*, *P. ovale*, and *P. vivax* from the high resolution (2592 px by 1944 px), but low quantity (210 images) public access malaria dataset (Malaria Parasite Image, Different Malaria Species). All images from this dataset are malaria positive. Arrows indicate examples of red blood cells infected with malaria.

identified images in order to "learn" how to classify unknown images in the future. In order to accurately make future diagnoses, machine learning must be set up to analyze and learn from available data. The model requires training data for the learning phase for the machine learning technique, which must be labeled by humans and then reformatted to the specifications of the specific machine learning model.

In order to conduct categorical classification with images, we used a convolusional neural network model. Neural networks are state-of-the-art machine learning algorithms that particularly excel in image analysis and recognition. Neural networks consist of interconnected neuron layers that pass data through several internal (hidden) layers, doing calculations along the way, and finally to an output set of neurons. The particular model we used was Google's InceptionV3 Convolutional Neural Network. The model is the third version of Google's Inception neural networks and was specially engineered to be trained on ImageNet's Large Visual Recognition Challenge, with established applications of this model in cancer research (12).

We hypothesized that machine learning involving the training of the model on a different dataset could avoid many of the problems with traditional machine learning for diagnosing malaria, specifically the requirement of large amounts of training data, which often is not available. This technique of machine learning, called transfer learning, could be inherently generalizable and cost-effective, particularly given its success in similar applications. Transfer learning is a methodology in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks. Transfer learning has additionally been used in several medical areas, such as the identification of COVID-19 in X-ray images (13).

Using transfer learning is identical to normal machine learning, except with a baseline for all the weights in the model (**Figure 1**). Before a pre-trained network can be used to analyze another dataset, the last row of neurons (known as the output layer) must be modified so that the correct answer can be generated through that layer. For example, if the original training dataset has one binary output and the new dataset has four choices (A, B, C, and D), the last layer must be changed from one to four nodes so that the neural network can actually choose between all four outputs. One of the most popular transfer learning datasets is ImageNet, a database of over 1 million labelled images that are from multiple sources. The abundance of these images as well as the wide variety of image types provide ample pre-training for neural networks, so we expected that ImageNet would be an excellent choice to use for transfer learning.

We identified a dataset from a machine learning dataset source, Kaggle, called "Malaria Parasite Image (Different Malaria Species)" (10), with limited examples showing the multiple different types of malaria in a classification dataset (**Figure 2**). This dataset has four separate categories of high-quality images of blood slides corresponding with four different Plasmodium species: *P. falciparum*, *P. malariae*, *P. ovale*, and *P. vivax*. The challenge with the dataset, however, is that it had 210 blood slides, which is considered a low number of examples. Thus, training a machine learning algorithm or neural network has the potential to be inaccurate

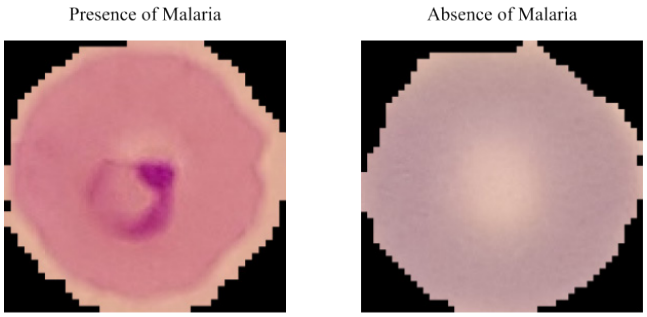

Figure 3: Sample blood slides from the low-resolution malaria dataset.
Examples of malaria blood slides from the low resolution, but higher quantity (27,558 images) public access malaria dataset (Malaria Cell Images Dataset). This binary classification dataset has both infected (left) and uninfected (right) categories.

due to sampling error.

Our primary hypothesis was that using transfer learning via ImageNet would result in increased accuracy of the model's predictions by the trained network rather than training and testing solely on the aforementioned small classification dataset. However, we also wanted to ascertain whether a more relevant dataset would be more effective in correctly diagnosing than ImageNet alone. Our thinking was that pre-training a network on data from a related problem may outperform a general image database like ImageNet, since the data would be more relevant to the topic being tested, resulting in higher accuracy. Thus, we identified a more` relevant dataset called "Malaria Cell Images Dataset" (11), which is a malaria binary classification dataset from Kaggle, consisting of 27,558 low resolution images of blood slides. Roughly half of the images are marked as infected while the other half are marked as uninfected (**Figure 3**). This "relevant dataset" is large enough in comparison to the complexity of the machine learning task at hand, so we predicted that our algorithms would have enough training data to show greater accuracy than transfer learning from ImageNet alone (**Table 1**).

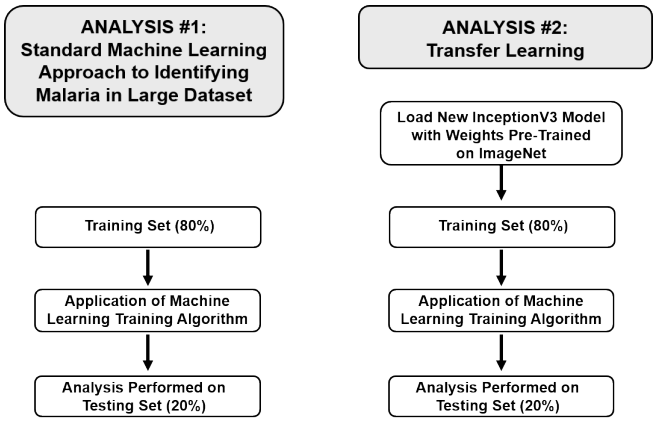

Figure 4: Overview of experimental plan, starting with traditional machine learning and then adding the transfer learning technique.
Of note, additional transfer learning analyses were performed when pre-training the algorithm using the large malaria dataset for application to the small malaria dataset (both with and without additional pre-training using InceptionV3), but these were unsuccessful due to differences in data quality and data categories between the two datasets.

Characteristic	Small Malaria Dataset	Large Malaria Dataset
Source	kaggle.com	kaggle.com
Number of Slides	210	27,558
Image Size	Fixed	Variable
Image Resolution	~2592 x 1944 pixels	Averaging 130 x 130 pixels
Control or Negative Slides	No	50% of data
Data Structure	Categorical	Binary

Table 1: Comparison of the two malaria datasets.

RESULTS

We performed tests to determine if using transfer learning with ImageNet as a source would improve accuracy over traditional machine learning. We applied the same InceptionV3 base model and settings for both the smaller and larger malaria datasets, both with and without transfer learning. The metrics we were testing for were validation accuracy, precision, recall, and F_1 score. Validation accuracy is the likelihood that the model correctly categorizes the given image (number of correct classifications divided by the total number of classifications performed). Recall and precision are machine learning equivalents of sensitivity and positive predictive value, respectively. The F_1 score is an overall score given to the results considering both recall and precision.

We trained the InceptionV3 model with no preloaded weights on the 80% randomly-selected training subset of the smaller malaria dataset, then applied the model to the 20% testing subset (**Figure 4**). This yielded an overall validation accuracy of 65.5%. Then, we trained and tested another InceptionV3 model with ImageNet weights preloaded, and when applied to the same small malaria dataset, the validation accuracy improved to 80.7%, an absolute increase of 15.2% (**Table 2**). Next, we repeated this process of training and testing the larger malaria dataset. The test without transfer learning reached an accuracy of 95.5% (expected to be much higher than the small malaria dataset, given the much larger sample size of blood slides). When pre-training the model with ImageNet weights preloaded, the accuracy improved by 2.5%, to 98.0% overall validation accuracy for identifying malaria on the slides. Similar modest improvements were seen when evaluating other statistical metrics (precision, recall, F_1 score)

		Validation Accuracy*	Precision (PPV)	Recall (sensitivity)	F_1 Score
Small Malaria Dataset	Without Transfer Learning	65.5%			
	With Transfer Learning	80.7%			
Large Malaria Dataset	Without Transfer Learning	95.5%	83.3%	91.1%	87.0%
	With Transfer Learning	98.0%	85.8%	92.1%	88.8%

Table 2: Applying transfer learning to the diagnostic algorithm.
Improvement in diagnostic accuracy after ImageNet pre-training of the InceptionV3 neural network model on the smaller and larger datasets of malaria blood slides. PPV = Positive Predictive Value. *Accuracy of the model on the validation subset of data.

on the large malaria dataset (**Table 2**). Interestingly, when training the InceptionV3 model on the larger malaria dataset alone, and then transferring the model to the smaller dataset (i.e., without any ImageNet weights preloaded), there was only a slight increase in accuracy (50.1% to 50.8%).

Due to the marginal improvement found by applying transfer learning with the large malaria dataset to the small malaria dataset, we decided to perform an exploratory analysis to determine if transfer learning would be more statistically relevant with smaller sample sizes. When reducing the training data to 50%, 25%, 12.5%, and 6.25% of the larger malaria dataset's original size, we found progressively greater impact of transfer learning as the sample size decreased on the validation accuracy of the final model (**Table 3**). This suggests that smaller datasets have more to benefit from transfer learning than larger ones do.

DISCUSSION

In this pilot study of transfer learning approaches to diagnosing malaria, we found that transfer learning, involving the training of a machine learning algorithm on a previous dataset, was a feasible approach to diagnosing malaria through computer-based technology. Applying transfer learning resulted in a substantial boost in accuracy when applied to a small sample size, improving accuracy by more than 15% in the smaller malaria dataset, with lesser improvement when starting with a larger dataset at higher baseline accuracy which had about a 2.5% absolute improvement. Of note, and contrary to what we predicted, transfer learning from a related dataset was less effective than using the very large nonspecific ImageNet database, even though the relevant malaria dataset was substantially more closely related to the final testing data.

While considering this approach, we recognized the limitations of basic machine learning algorithms with images, particularly when lower-quality microscope slides would be the source of all data. The greater success of ImageNet in transfer learning over the related malaria images is likely related to ImageNet's enormous pool of images (over 1,000,000) versus the relevant malaria dataset (only 27,558 images). This is an important interpretation since it implies the greater importance of repetition and volume over individual detail when using transfer learning to pre-train machine learning algorithms for diagnosing disease. Alternatively, it is possible that transfer learning from a dataset with a different structure (i.e., transferring from the binary classification dataset to the 4-category small dataset, or vice versa) could cause a significant loss of training power, and thus accuracy. Again, this has important implications when attempting to apply machine learning to diseases like malaria, in which multiple sub-classifications may be present in any given dataset or any given patient being tested.

A key strength of machine learning applications is the ability to rapidly deploy this type of application in the real world, as a mobile and easily attainable solution if developed for distribution. One current example is a smartphone application for Android with a lens attachment that mounts on the phone's built-in camera to read a microscope slide in increased detail (9). This app uses a pre-trained machine learning algorithm to make a diagnosis automatically, so no specific training or skilled personnel are needed beyond the creation of the blood slide from an individual patient. This magnification app would be made available on the Android store, which would be significantly less expensive and more accessible in remote areas when compared with standard light microscopy or RDTs (9). Thus, it is paramount to advance the performance of such applications that are becoming more easily accessible in areas where other approaches are not feasible.

Of note, the process of collecting data requires extra work initially, as researchers or clinicians must save images of blood slides, classify them, verify their accuracy, and compile them into a dataset. For this reason, data collection may not

Portion of Large Malaria Dataset	With/Without Transfer Learning (TL)	Validation Accuracy*	Precision (PPV)	Recall (sensitivity)	F_1 Score
100%	Without TL	94.4%	89.5%	87.7%	87.6%
	With TL	95.1%	90.0%	89.6%	88.9%
50%	Without TL	94.1%	86.8%	91.8%	88.3%
	With TL	95.7%	90.1%	92.0%	90.3%
25%	Without TL	90.0%	81.9%	91.4%	85.0%
	With TL	95.5%	89.4%	90.1%	89.0%
12.5%	Without TL	69.5%	61.4%	41.4%	47.4%
	With TL	95.3%	90.4%	90.4%	89.6%
6.25%	Without TL	51.7%	51.6%	95.3%	64.4%
	With TL	91.9%	90.1%	84.7%	86.6%

Table 3: Effect of transfer learning on progressively smaller sample size.
PPV = Positive Predictive Value. *Accuracy of the model on the validation subset of data

be possible in the remote regions of Africa where the disease is most active. However, data can be analyzed in hospitals and testing centers that are located in more developed regions, to avoid excess workload in rural centers. This would allow the remote health centers to upload data which can then be compiled so that an accurate dataset can be presented to the algorithm to learn. After the algorithm or neural network has been adjusted to fit the data, it can then be applied in more remote areas where the quality of malaria testing is poor.

There are several important limitations to consider with this research study. First, the low accuracy resulting from the transfer learning from the larger malaria dataset could be addressed using several techniques in the future to potentially improve results. One option is to find a new dataset, as the two malaria datasets tested in our study differ from each other considerably in terms of quality and output type (i.e., binary versus 4 categorical results). In addition, the larger dataset we used for pre-training has lower quality images that may be too dissimilar to the higher-quality images found in the smaller dataset. Another potential approach could be to use unsupervised learning techniques like K-means clustering, a form of machine learning that minimizes the need for specific labelling of the data beforehand. This would allow the machine learning algorithm to specify a number of distinct groups within the dataset (designated as "K" number of groups), then using the K-means clustering approach in higher power computers. The program could then group cells into infected and uninfected regions (8). A second important limitation is the inability to distinguish between the multiple different species of malaria, largely due to the small number of examples available on public databases of malarial blood slides. Future studies could help train the machine learning algorithms on larger and more robust datasets, to help direct diagnosis and treatment of the multiple Plasmodium subtypes.

Although this was a limited pilot study, our application of transfer learning demonstrates the potential utility of this approach to diagnosing disease using existing databases and technologies, including those readily available on a smartphone. Future studies should evaluate whether refinement of this approach may help improve diagnostic accuracy, and thus allow regions of the world with limited availability of medical care to provide appropriate therapies to their populations.

MATERIALS AND METHODS

During the research we exclusively used Python and imported Tensorflow and Keras to assist with the machine learning. For each series of tests, we loaded a version of Keras's built-in InceptionV3 neural network with either no preloaded weights or with preconfigured weights trained on ImageNet. After the network was created, we added a global average pooling layer and a dense layer of 1024 nodes, plus a custom output layer to match what each dataset needed for output. Specifically, the binary classification dataset required

one output node while the categorical dataset required four: *P. falciparum*, *P. malariae*, *P. ovale*, and *P. vivax*. This approach—using an existing neural network and modifying the output for our specific purposes and number of outputs required—is a standard method used in transfer learning research. We then imported all of the images, resized them to a constant size, and split them randomly into train and test data using Keras's ImageDataGenerator method. All images needed to be resized to 299 pixels by 299 pixels, as that was the size of the input layer of the InceptionV3 network. For all tests we used an 80% to 20% train-test split. We then optimized the model using root means square (RMS) with a learning rate of $\lambda = 0.0001$. After compiling and optimizing the model, we ran the training with the training data (80%) and evaluated the model with the testing data (20%). We then printed out the metrics: validation accuracy, precision, recall, and F_1 score. Note that not all of the tests apply transfer learning (TL).

Small malaria without ImageNet (no TL)

We started the InceptionV3 network loaded with no weights (not pre-trained) and proceeded to train on 80% of the small dataset and then test on the remaining 20% of the samples in the small dataset. This was the control test of the small malaria dataset without any application of transfer learning.

Small malaria with ImageNet (TL)

We started the InceptionV3 network loaded with ImageNet weights (pre-trained with ImageNet) and proceeded to train on 80% of the small dataset and then test on the remaining 20% of the samples in the small dataset. This test is the same as the previous but includes transfer learning from ImageNet to the small malaria dataset.

Large malaria without ImageNet (no TL)

We started the InceptionV3 network loaded with no weights (not pre-trained) and proceeded to train on 80% of the large dataset and then test on the remaining 20% of the samples in the large dataset. This was the control test of the large malaria dataset without any application of transfer learning.

Large malaria with ImageNet (TL)

We started the InceptionV3 network loaded with ImageNet weights (pre-trained with ImageNet) and proceeded to train on 80% of the large dataset and then test on the remaining 20% of the samples. This test is the same as the previous but includes transfer learning from ImageNet to the large malaria dataset.

Other tests (both malaria with/without ImageNet) (TL)

We started the InceptionV3 network loaded with no weights and proceeded to train on the entire large dataset followed by 80% of the small dataset and then test on the remaining 20% of the samples in the small dataset.

We started the InceptionV3 network loaded with ImageNet weights and proceeded to train on the entire large dataset followed by 80% of the small dataset and then test on the remaining 20% of the samples in the small dataset.

Received: December 15, 2020
Accepted: September 13, 2021
Published: November 07, 2021

REFERENCES

1. World Health Organization. "Fact Sheet About Malaria." *World Health Organization*, 14 Jan. 2020, www.who.int/news-room/fact-sheets/detail/malaria.
2. McKenzie, F E, and W H Bossert. "Multispecies Plasmodium infections of humans." *The Journal of Parasitology*, vol. 85, 1999, pp. 12-18.
3. World Health Organization. "Malaria." *World Health Organization*, www.who.int/data/gho/data/themes/malaria.
4. ACTwatch Group., Akulayi, L., Alum, A. *et al*. "Testing times: trends in availability, price, and market share of malaria diagnostics in the public and private healthcare sector across eight sub-Saharan African countries from 2009 to 2015." *Malaria Journal*, vol. 16, 2017, p. 205. doi.org/10.1186/s12936-017-1829-5
5. World Health Organization. "How Malaria RDTs Work." *World Health Organization*, www.who.int/malaria/areas/diagnosis/rapid-diagnostic-tests/about-rdt/en/.
6. Hoffman, S. L., C. Campbell, N. White, *et al*. "Malaria." *Tropical Infectious Diseases*, vol. 2, 2006, pp. 1024-1062, ScienceDirect. doi.org/10.1016/B978-0-443-06668-9.50095-8.
7. Linder, N., R. Turkki, M. Walliander, *et al*. "A Malaria Diagnostic Tool Based on Computer Vision Screening and Visualization of Plasmodium falciparum Candidate Areas in Digitized Blood Smears." *PLoS ONE*, vol. 9, no. 8, 2014, p. e104855, doi.org/10.1371/journal.pone.0104855.
8. Nasir, A., M. Mashor, Z. Mohamed, *et al*. "Segmentation based approach for detection of malaria parasites using moving k-means clustering." *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, 2012, pp. 653-658, doi.org/10.1109/IECBES.2012.6498073.
9. Poostchi, M., K. Silamut, R. J. Maude, *et al*. "Image analysis and machine learning for detecting malaria." *Translational Research*, vol. 194, Apr. 2018, pp. 36-55, doi.org/10.1016/j.trsl.2017.12.004.
10. Saif Uddin. "Malaria Parasite Image (Different Malaria Species)." Kaggle: Your Machine Learning and Data Science Community, www.kaggle.com/saife245/malaria-parasite-image-malaria-species?select=DatasetDescription.pdf.
11. Arunava. "Malaria Cell Images Dataset." Kaggle: Your Machine Learning and Data Science Community, www.kaggle.com/iarunava/cell-images-for-detecting-malaria.
12. Intel. "Inception V3 Deep Convolutional Architecture For Classifying Acute Myeloid/Lymphoblastic Leukemia." Intel, software.intel.com/content/www/us/en/develop/articles/inception-v3-deep-convolutional-architecture-for-classifying-

acute-myeloidlymphoblastic.html.
13. Apostolopoulos, I. D., and T. A. Mpesiana. "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks." *Physical and Engineering Sciences in Medicine*, vol. 43, no. 3, 2020, pp. 635–640, doi.org/10.1007/s13246-020-00865-4.