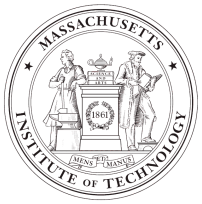
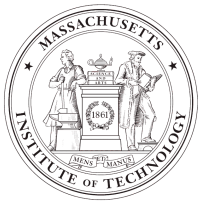


Massachusetts
Institute of
Technology

Capstone Project Machine Learning Final Report: Malaria Detection

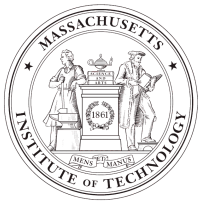
Daphne Pariser
February 2023t



Massachusetts Institute of Technology

Data Dictionary

RBC- Red blood cells
ANN- Artificial neural networks
CNN- Convolutional neural networks
PCR- Polymerase chain reaction
ML- Machine learning
AI- Artificial intelligence
NN- Neural networks
RGB- Red, Green, Blue
HSV- Hue, Saturation, Value



Massachusetts Institute of Technology

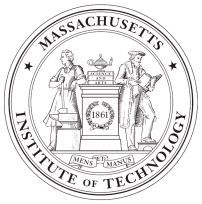
Executive Summary

Here I have proposed a machine learning algorithm that utilizes neural networks (NN) to classify images of parasitized or uninfected red blood cells Geimsa stained for malaria with greater accuracy. Traditionally, the detection of infected cells from a blood sample is performed through microbiological analysis by a microscope, followed by specialist interpretation of the results to arrive at a diagnosis. The dataset consists of 24,958 train and 2,600 test images of colored microscope samples stained with Geimsa to identify Plasmodium parasite infection. Using this automated machine learning and AI approach to diagnosing infection has many advantages, including reducing human error, increasing efficiency, and increasing accuracy. By incorporating deep learning techniques from the computer vision field, the proposed framework presents deep learning architectures based on Convolutional Neural Networks (CNN) to detect malaria-infected cells accurately. The CNN model designed in this work achieves a 99% accuracy rate in identifying malaria-infected red blood cells. Ultimately, the developed CNN model has significant outcomes for profitability, including efficient, accurate, and automated malaria diagnosis, reduced human error, and improved accuracy in interpretation, leading to increased customer satisfaction and cost savings long term. Additionally, with simple modifications, this could be utilized for disease outbreak tracking and monetized on a more global scale.

Background/Problem Statement

The main problem is that malaria is a life-threatening disease caused by the *Plasmodium* parasite and is transmitted to people through a mosquito bite and is geographically predominant in Africa, Latin America, and Asia. Malaria is the leading cause of death in areas where malaria is prevalent, with the most vulnerable groups being young children and pregnant women. Malaria is a single-cellular organism (1) that first travel to the liver, where they infect hepatocytes; this is where the parasite grows, divides, and matures over the course of ~10 days, and at maturity, one hepatocyte can hold up to 40,000 parasites (2). The infected hepatocytes eventually lyse, and the parasites are released into the bloodstream, interacting with red blood cells (RBCs) and infecting them. The parasite continues to divide over the next 24-72 hours after infecting an RBC. Then the RBC lyses cause more parasites in the bloodstream, hemolysis, and eventually anemia in malaria-infected patients (2). There are no symptoms when individuals have liver infections, symptoms occur only at the stage of RBC infection, and once there are 100,000 parasites per milliliter of blood (1). According to the WHO, in 2021, there were over 247 million malaria cases worldwide, and over 619,000 people died from malaria that year.

Non-governmental organizations, for-profits, and governments have worked together for several decades to create interventions, preventative measures, cures, and vaccinations for malaria. From 2010 to 2020, we have seen a significant decrease in the overall mortality of malaria worldwide by as much as 36% (3). There are four malaria parasites *Plasmodium falciparum, P.*



Massachusetts Institute of Technology

vivax, P. ovale, and *P. malariae*. Recently, a vaccine has been developed against *Plasmodium* falciparum, the malaria parasite known to cause the most severe disease, especially in children, and can cause cerebral malaria (4). The efficacy of this vaccine is still being evaluated; additionally, vaccines for other *Plasmodium* parasites still need to be created. Therefore, having technology capable of machine learning algorithms that allow for fast-paced detection methodologies of parasitized red blood cells could aid in the overall global prevention, detection, and monitoring of vulnerable populations.

Additionally, machine learning algorithms can aid with the reduction of resource-heavy laboratories and the need for excessive funding to diagnose patients accurately. The stain used in this particular methodology is called a Geimsa stain (5) and is the standard methodology worldwide for determining *Plasmodium* infections. Several methods for detecting malaria in blood samples include a microscopy examination of a Geimsa-stained RBC slide and polymerase chain reaction (PCR) to detect the parasite's DNA. PCR is not often used due to the high cost of the technology needed and the cost of education and technical expertise. Both these methodologies are time-consuming for laboratories, and appropriately trained technicians are not always available in certain resource-limited areas. The utilization of Geimsa stain allows laboratories with limited resources to detect plasmodium, infected patients, and the use of a machine learning and artificial intelligence (ML/AI) algorithm can aid in the accuracy and speed at which a patient can be diagnosed. The algorithm would also allow laboratories without technicians to submit photos and diagnose patients; this would be exceptionally useful for laboratories in resource-limited areas that are unavailable for technicians or high-level equipment. Although the PCR methodologies are much more sensitive and accurate, using ML/AI algorithms can help develop more precise detection and diagnostic methods for microscopy examination.

Design Solution Summary

An automated CNN was implemented for the diagnosis of malaria using 24,958 train and 2,600 test images of colored microscope samples stained with Geimsa to identify Plasmodium parasite infection. In making the binary-class classification of the parasitized and uninfected cells in this paper, I consider the image size input of 64 x 64 pixels. The CNN consists of dense layers with 512 nodes, activated by the Softmax function. The Adam optimization algorithm is used at a learning rate of 0.0001. Figure 1 illustrates the Convolutional Neural Network Architecture, which classifies parasitized and uninfected red blood cells. Artificial neural networks were used to categorize images of parasitized and uninfected cells (Figure 2). The configuration of the CNN is outlined in Table 2.

Using an artificial neural network, we are able to identify that the images were evenly distributed among both the training and test data, where we find that there are 12,582 parasitized images

and 12,376 uninfected images, and 1,300 test images for both the parasitized and uninfected cohort (Table 1). This balance is crucial in training machine learning models as it helps to avoid any biases. Models trained on imbalanced data are more likely to have a skewed understanding of the problem, resulting in inaccurate predictions. For instance, if a model is trained on a dataset heavily skewed towards one class, it may constantly predict that class even when presented with instances of the other class, resulting in poor performance and unreliable outcomes. However, a balanced dataset offers the model a comprehensive representation of the problem, enabling it better to learn the patterns and relationships in the data and make more precise predictions.

The code incorporated here utilized image augmentation, which brings several advantages, including a larger training dataset. Image augmentation artificially expands the training data, allowing the model to generalize better to new images by exposing it to more diverse variations of the same objects and scenes. This technique can be especially beneficial when the original training data is limited in size or diversity and helps the model avoid overfitting, leading to improved validation set performance. The image augmentation utilized here was horizontal flip, zoom range, rotational range, and fill mode set to constant.

By flipping the images horizontally (using a horizontal flip) and using rotational range, the model is exposed to new variations of the same objects, scenes, and patterns, which can help it generalize better to new images. This is especially useful when the original training dataset is small or lacks sufficient diversity. It helps the model avoid overfitting and can improve performance on the validation set. The horizontal flip augmentation can help the model become more robust to various orientations of the objects in the images and improve its ability to detect parasitized and uninfected red blood cells accurately.

The zoom augmentation can help the model accurately detect parasitized and uninfected red blood cells. For example, the model may encounter images with blood cells that are slightly smaller or larger than those in the training data, and the zoom augmentation can help the model become more robust to these variations. In this instance, I found it was more useful to have a larger zoom to ensure the entire cell was captured and the entirety of the parasite was then gauged.

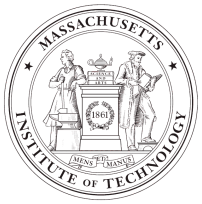
The use of the fill mode set to constant in image augmentation is useful because it helps to preserve the background information and maintain the original aspect ratio of the image. When images are rotated or transformed during augmentation, new pixels are created, and the fill mode determines what values should be used to fill these pixels. By setting the fill mode to "constant", the new pixels are filled with a constant value, which helps to preserve the background information and maintain the original aspect ratio of the image. This is especially useful in preserving the context of the image and maintaining the continuity of the background. By preserving the background information, the model is able to better understand the

relationship between the objects of interest (e.g. parasitized and uninfected red blood cells) and their surrounding context, which can improve its ability to make accurate predictions.

Although in my second report the image augmentation did not improve the model much, I was able to add more convolutional layers and improve my model, obtaining a higher accuracy. The combination of image augmentation with leakyReLU and batch normalization resulted in test accuracy, recall, precision, and F1-score all around 99%, however, it also showed some signs of underfitting. Hence, increasing the amount of training data could be a useful solution. Despite this, the model still performs exceptionally well and can be used for malaria detection. In the following sections, ways to improve the model will be explored.

The model described above is the best model created, but there are several models that were designed that looked at numerous iterations of CNNs with and without image augmentation. In comparing these models test accuracy, recall, precision, and F1-score were all considered as well as the computational cost of running these models (Table 3). As seen in Table 3 there are similarities in the recall, precision, and F1-score between many of the models but the computational cost to run these models varies significantly, putting model 3 and model 5 as the most computationally efficient models. When comparing model 3 and model 5 underfitting was also considered and as can be seen in Figure 4 model 5 had the least amount of underfitting compared to its model 3 counterpart.

When comparing Model 5 and Model 3, there are several factors that were taken into to consideration. The accuracy of Model 5 is 0.984 while Model 3 has an accuracy of 0.983. This indicates that Model 5 is more accurate than Model 3. The running time/sample of Model 5 is 3 ms per sample while Model 3 has a running time of 2.92 ms per sample. Model 5 is slightly slower, but the difference is negligible. The precision of both models is 0.98 for uninfected and 0.99 for parasitized, however, Model 5 has a higher precision for uninfected and a lower precision for parasitized, which indicates that Model 5 is better at identifying uninfected cases. The recall for Model 5 is 0.99 for uninfected and 0.98 for parasitized, while Model 3 has a recall of 1 for uninfected and 0.97 for parasitized. This means that Model 5 is better at identifying all of the uninfected cases and almost as good at identifying all of the parasitized cases. The F1-Score for Model 5 is 0.99 for both uninfected and parasitized, while Model 3 has a F1-Score of 0.98 for both. This indicates that Model 5 is a better overall performer. Based on these factors, it can be argued that Model 5 is a better choice over Model 3 as it has a higher accuracy, a higher precision for uninfected, a better recall for uninfected, and a higher F1-Score overall, which provides the validity for why this model is the most efficient and useful model. The slightly slower running time is not a significant concern.



Massachusetts Institute of Technology

As we can see in our confusion matrices, both models perform exceptionally well, but there is always room for improvement (Figure 5). Confusion matrices are helpful in evaluating the performance of a machine learning model by summarizing the results of a classification problem. The matrix provides a clear and concise way to represent the number of correct and incorrect predictions made by the model. This allows for a comprehensive analysis of the model's performance and helps identify areas where the model may need improvement. Confusion matrices also help to understand the precision, recall, and F1-score of the model, which are important metrics for evaluating a model's accuracy. Additionally, confusion matrices can be used to determine the balance between false positive and false negative predictions, which can be particularly useful in real-world applications where false positives and false negatives may have different implications.

Thus there are some considerations to make for more accurate detection of malaria that should be considered:

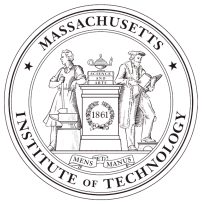
- Large and diverse dataset of images of infected and non-infected blood cells, it may be that the training dataset we currently are working with is too small. Online databases could be a good way to increase the training datasets
- Robust feature extraction for extracting important features from images that can effectively differentiate between infected and uninfected cells and here we might consider utilizing several features together to create a stronger model and accurately identify and diagnose malaria.
- Effective hyperparameter tuning considerations involving experimenting with different combinations of hyperparameters, such as learning rate, number of hidden layers, and batch size, to find the optimal set of hyperparameters that results in the best performance on the model's evaluation metric. Methods to consider are grid search, random search, or Bayesian optimization.

Thus there is scope to further improve this model, by focusing on underfitting, feature extraction, hyperparameter tuning, and increasing the training dataset.

Recommendations & future directions

The literature is flourishing with scientific research relating to the detection of parasitized and uninfected by malaria often utilizing national databases as training data and various deep learning approaches. The work presented here contributes to the existing literature by analyzing a convolutional neural network designed with image augmentation. The benefit of using the CNN model described that utilizes Giemsa-stained images to detect parasitized and uninfected patients is that it can potentially provide a more accurate and efficient way to diagnose malaria.

There are several key stakeholders for this CNN model. Healthcare organizations can implement the CNN model as an aid in the diagnostic process and ensure that patients are



Massachusetts
Institute of
Technology

appropriately diagnosed and treated. Researchers could help to collaborate with other organizations to increase the scope of research and improve our understanding of malaria. AI developers could continuously monitor and evaluate the performance of the neural network model. Government and policymakers can help to ensure that the technology is used ethically and responsibly by considering ethical implications and protecting patient privacy.

In affected countries, the general public can also be considered as stakeholders with this CNN model. They can benefit from the improved accuracy and efficiency of malaria diagnosis, reducing the spread of the disease and ultimately improving public health. Patients who are diagnosed with malaria could receive faster and more accurate treatment, leading to improved outcomes and reduced risk of complications. Additionally, local healthcare providers and community health workers can play an important role as stakeholders in implementing and using the CNN model. They can provide input and feedback on the feasibility and acceptability of the technology in their communities and help to address any barriers to implementation.

Furthermore, stakeholders such as non-government organizations, international aid organizations, and philanthropic organizations can help provide funding and resources for implementing and disseminating the CNN model in affected countries. They can also help to promote the technology to healthcare providers and ensure that it is accessible and affordable for patients in need.

In terms of consideration for for-profit models, subscription models can be considered as malaria affects people seasonally. Additionally, in collaboration with other partners, this can be used for outbreak monitoring. Lastly, with slight modifications, this can be utilized for other diseases that require binary classification, such as COVID-19.

Key recommendation for the implementation of this CNN is to consider partnering with other organizations that utilize paper-based assays to include more patient data in this neural network. By incorporating data from these assays, we can gain a more holistic understanding of the patient's health, including information such as their complete blood count, plasma, and other factors that may impact the progression of their malaria infection. This can help ensure that patients are appropriately diagnosed and treated, which can significantly reduce the overall morbidity and mortality associated with malaria.

By including this additional data, we can more accurately track patients over time, which can help to identify patterns and trends that may be missed by relying on a single diagnostic method. This could be particularly useful in resource-limited settings where follow-up visits may be infrequent, and accurate tracking of patient health is essential for providing proper care.

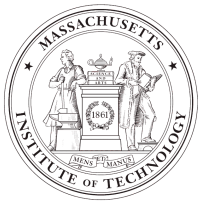
Finally, by working with other organizations, we can also increase the scope of the research, which can help to further our understanding of malaria and improve our ability to diagnose and

treat it. This can be a key step to improving patient care and reducing the global burden of malaria.

Other recommendations should also consider ensuring that the technology is used ethically and responsibly by considering ethical implications and protecting patient privacy, continuously monitoring and evaluating the performance of the neural network model to ensure it remains accurate and efficient and partnering with healthcare organizations to provide appropriate care for patients diagnosed with malaria.

The costs associated with this project include an initial investment in developing and implementing this CNN model. There would also be ongoing costs for maintaining and updating the model to ensure its accuracy and efficiency. In addition, the computational cost of running these models must also be considered. Model 5 described is the best one designed, with several other models designed and compared based on numerous iterations of CNNs with and without image augmentation. The results indicate that Model 5 is the most computationally efficient model, with a slightly slower running time of 3 ms per sample compared to Model 3's running time of 2.92 ms per sample, but as stated previously, Model 5 is also the better performer, with higher accuracy, precision for uninfected, recall for uninfected, and a higher F1-Score overall. When considering the cost associated it is important to consider how the training data was collected to ensure that there was proper training of the CNN. We already have national databases where images are sent in and thus we should utilize this as a way to create a standard model the way that the human genome project was done.

There are several key risks and challenges associated with this project, including ensuring the ethical and responsible use of technology by taking into account ethical implications and protecting patient privacy. Another challenge is maintaining the accuracy and efficiency of the neural network model, which requires ongoing monitoring and evaluation. Additionally, incorporating data from other organizations that use paper-based assays and, generally, more training data to provide a more comprehensive understanding of patients' health is also a challenge that needs to be addressed. These risks and challenges must be effectively managed to ensure the project's successful implementation and long-term sustainability. The proposed solution for model 5 has several challenges that could impact its accuracy. The main issue is that the model's ability to recognize new images could be limited if the training data used to build the model was limited in size and diversity. This could cause the model to overfit, leading to poor results and unreliable outcomes. Additionally, the model may not be able to handle different orientations of objects in the images and may not accurately identify infected red blood cells. The model may also struggle to identify infected cells if the images in the test data are slightly different in size from those in the training data. If the background information is not preserved during image processing, the model may not understand the relationship between the objects of interest and their surroundings. Finally, the computational cost of running the model may vary greatly, affecting its practicality and use in real-world applications. The model may also



Massachusetts
Institute of
Technology

not perform well if it does not have enough training data and may need more data to improve its accuracy.

In conclusion, using a CNN model designed with image augmentation to diagnose malaria can provide a more accurate and efficient method for detecting parasitized and uninfected patients. This technology has the potential to benefit healthcare organizations, researchers, AI developers, government and policymakers, and the general public, including patients, local healthcare providers, and community health workers. The implementation of this CNN model could greatly improve the overall accuracy and efficiency of malaria diagnosis, reducing the spread of the disease and improving public health. In summary, the implementation of this CNN model for malaria diagnosis can bring significant benefits and improve the overall public health outcomes. This project has the potential to make a significant impact in the global fight against malaria and improve the quality of life for those affected by the disease.

Figures

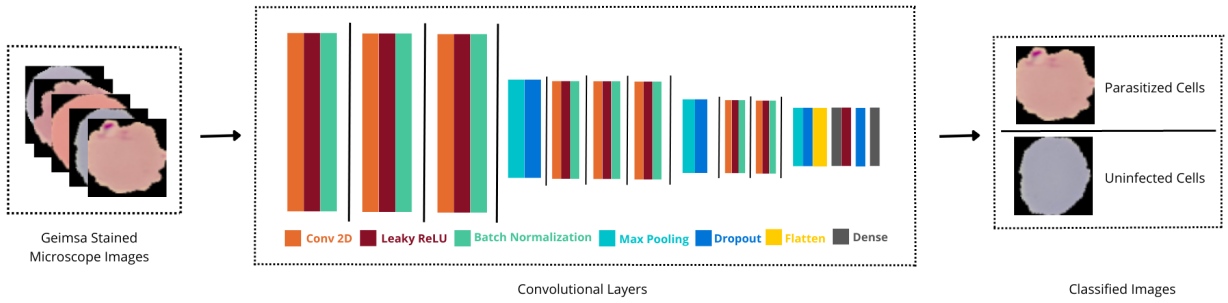


Figure 1. A dataset is provided to train the CNN model where convolutional layers are used along with LeakyReLU, Batch Normalization, Max Pooling, Dropout, Flatten, and Dense to detect parasitized and non-infected red blood cells.

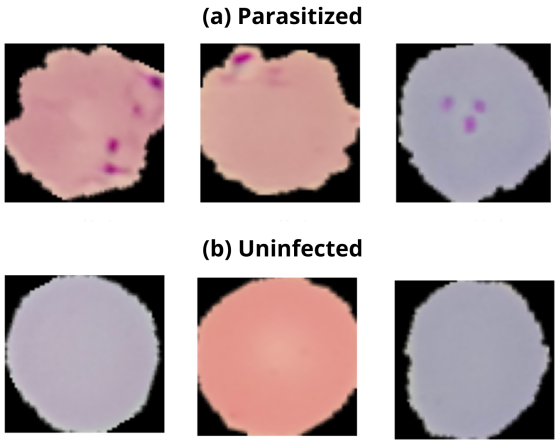


Figure 2. Example images from the dataset illustrating both (a) parasitized and (b) uninfected red blood cells.

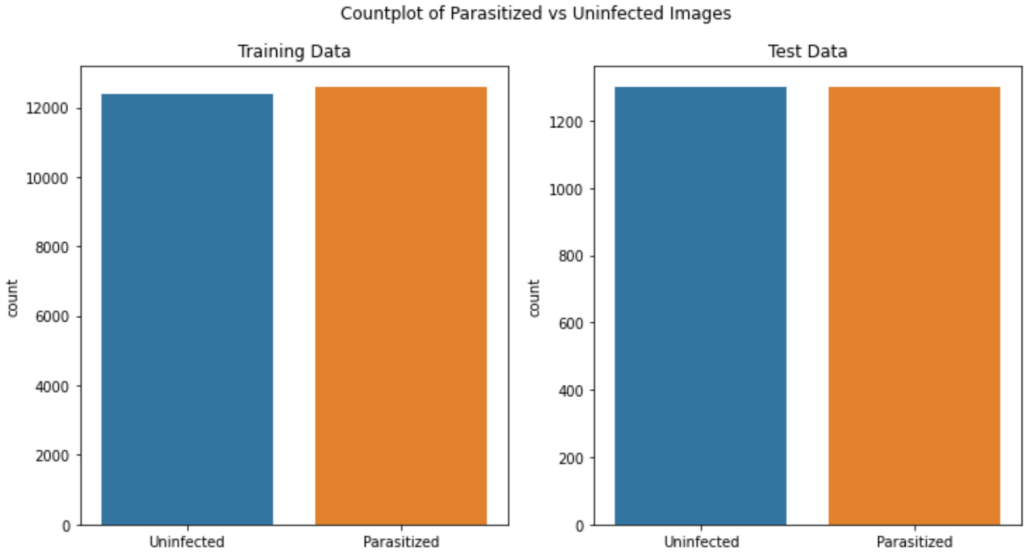


Figure 3. The training and test datasets provided are well-balanced.

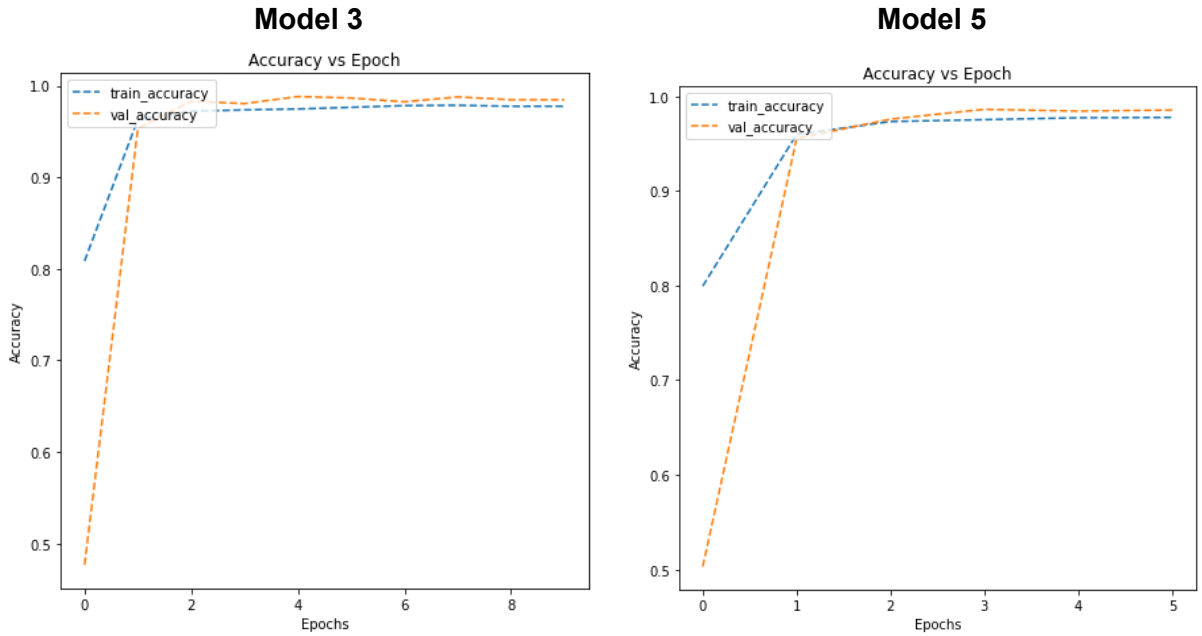


Figure 4. Compared to model 3, model 5 has less underfitting and therefore is likely the more efficient model at classifying data.

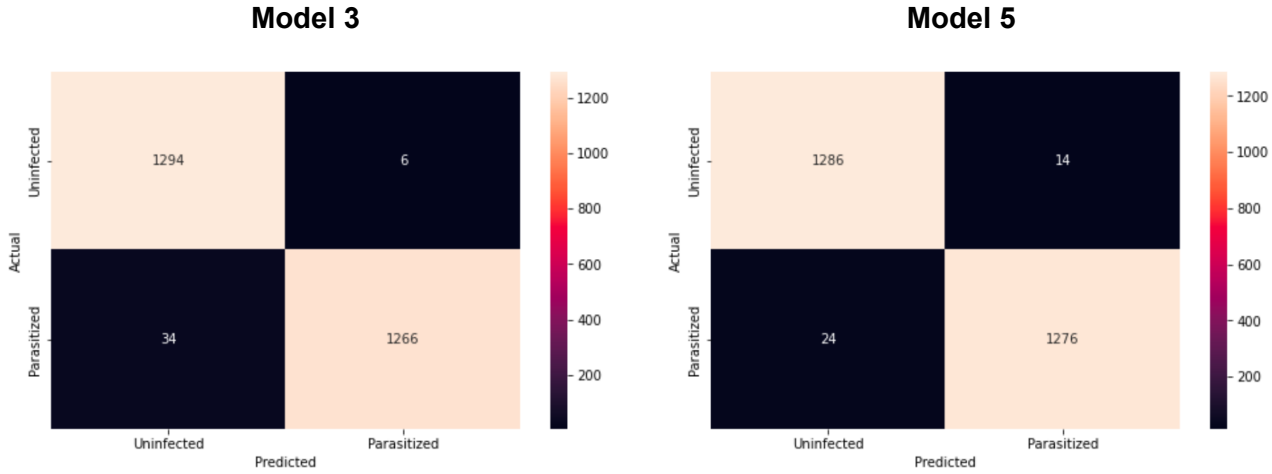
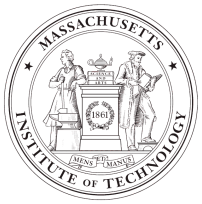


Figure 5. Confusion matrices of model 3 and model 5.



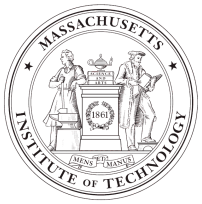
Massachusetts Institute of Technology

Count of Data		
Training Data	Parasitized	12,582
	Uninfected	12,376
Test Data	Parasitized	1,300
	Uninfected	1,300

Table 1. The count of training and test datasets shows that the dataset is well-balanced.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 64, 64, 32)	416
leaky_re_lu (LeakyReLU)	(None, 64, 64, 32)	0
batch_normalization (BatchNormalization)	(None, 64, 64, 32)	128
conv2d_1 (Conv2D)	(None, 64, 64, 32)	9248
leaky_re_lu_1 (LeakyReLU)	(None, 64, 64, 32)	0
batch_normalization_1 (BatchNormalization)	(None, 64, 64, 32)	128
conv2d_2 (Conv2D)	(None, 64, 64, 32)	9248
leaky_re_lu_2 (LeakyReLU)	(None, 64, 64, 32)	0
batch_normalization_2 (BatchNormalization)	(None, 64, 64, 32)	128



Massachusetts Institute of Technology

max_pooling2d (MaxPooling2D (None, 32, 32, 32) 0
)

dropout (Dropout) (None, 32, 32, 32) 0

conv2d_3 (Conv2D) (None, 32, 32, 64) 8256

leaky_re_lu_3 (LeakyReLU) (None, 32, 32, 64) 0

batch_normalization_3 (Batc (None, 32, 32, 64) 256
hNormalization)

conv2d_4 (Conv2D) (None, 32, 32, 64) 36928

leaky_re_lu_4 (LeakyReLU) (None, 32, 32, 64) 0

batch_normalization_4 (Batc (None, 32, 32, 64) 256
hNormalization)

conv2d_5 (Conv2D) (None, 32, 32, 64) 36928

leaky_re_lu_5 (LeakyReLU) (None, 32, 32, 64) 0

batch_normalization_5 (Batc (None, 32, 32, 64) 256
hNormalization)

max_pooling2d_1 (MaxPooling (None, 16, 16, 64) 0
2D)

dropout_1 (Dropout) (None, 16, 16, 64) 0

conv2d_6 (Conv2D) (None, 16, 16, 128) 73856

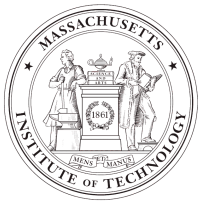
leaky_re_lu_6 (LeakyReLU) (None, 16, 16, 128) 0

batch_normalization_6 (Batc (None, 16, 16, 128) 512
hNormalization)

conv2d_7 (Conv2D) (None, 16, 16, 128) 147584

leaky_re_lu_7 (LeakyReLU) (None, 16, 16, 128) 0

batch_normalization_7 (Batc (None, 16, 16, 128) 512
hNormalization)

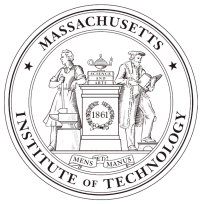


Massachusetts Institute of Technology

max_pooling2d_2 (MaxPooling 2D)	(None, 8, 8, 128)	0
dropout_2 (Dropout)	(None, 8, 8, 128)	0
flatten (Flatten)	(None, 8192)	0
dense (Dense)	(None, 512)	4194816
leaky_re_lu_8 (LeakyReLU)	(None, 512)	0
dropout_3 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 2)	1026

Total params: 4,520,482
Trainable params: 4,519,394
Non-trainable params: 1,088

Table 2. The architecture of Model 5.



Massachusetts Institute of Technology

Model	Infection	Accuracy	Running time/sample/sec	Precision	Recall	F1-Score
Model 1	Uninfected	0.983	~7.2 ms	0.99	0.98	0.98
	Parasitized			0.98	0.99	0.98
Model 2	Uninfected	0.985	~7.4 ms	0.99	0.98	0.98
	Parasitized			0.98	0.99	0.99
Model 3	Uninfected	0.984	~2.92 ms	0.97	1	0.98
	Parasitized			1	0.97	0.98
Model 4-VGG16	Uninfected	0.948	~21 ms	0.93	0.97	0.95
	Parasitized			0.97	0.93	0.95
Model 5	Uninfected	0.984	~3 ms	0.98	0.99	0.99
	Parasitized			0.99	0.98	0.99

Table 3. Comparison of models designed to classify parasitized and uninfected red blood cells.

Bibliography

1. Malaria [Internet]. [cited 2023 Jan 24]. Available from: https://www.who.int/en/news-room/fact-sheets/detail/malaria
2. Cowman AF, Healer J, Marapana D, Marsh K. Malaria: biology and disease. Cell. 2016 Oct 20;167(3):610–24.
3. CDC - Malaria - Malaria Worldwide - Impact of Malaria [Internet]. [cited 2023 Jan 23]. Available from: https://www.cdc.gov/malaria/malaria_worldwide/impact.html
4. Kihara M, Carter JA, Newton CRJC. The effect of Plasmodium falciparum on cognition: a systematic review. Trop Med Int Health. 2006 Apr;11(4):386–97.
5. Barcia JJ. The Giemsa stain: its history and applications. Int J Surg Pathol. 2007 Jul;15(3):292–6.