Executive Summary

In this project, a Convolutional Neural Network (CNN) model is proposed as a solution to the problem statement of building an accurate and quick computer vision model to detect malaria. After data visualization methods were used on the provided Malaria cell images dataset, it was seen that the Parasitized Cell contains a purple particle within the cell that shows the Plasmodium parasite responsible for causing Malaria. Uninfected cells do not contain that particle however, they may contain other particles that would need to be incorporated and learned by the computer vision model. A closer look using Hue Saturation Value (HSV) images allowed to see a better contrast of the colors which highlighted the parasite more prominently in the Parasitized cells' images. The proposed model was chosen to be a CNN as they are adept at learning localization within the image to identify if the parasite is present in the cell. Multiple models were trained using different architectures and datasets, including a model with augmented images and a transfer learning pre-trained model. However, the best trained Model had a recall of 0.97 for Uninfected cells and a recall of 0.99 for Parasitized cells. The overall accuracy of this model was the highest out of all six trained models at 98.77%. However, the model has certain limitations such as in order to use the model, images will need to be gathered in the same way as the training dataset images. The model may be improved by having an approach within the model that can normalize the input images so that they look the same. Due to this limitation, it is recommended that in order to use this model, stakeholders obtain images with the same dye, image resolution and image modality that used in images within the training dataset to have the model behave optimally.

Problem Summary and Solution Design

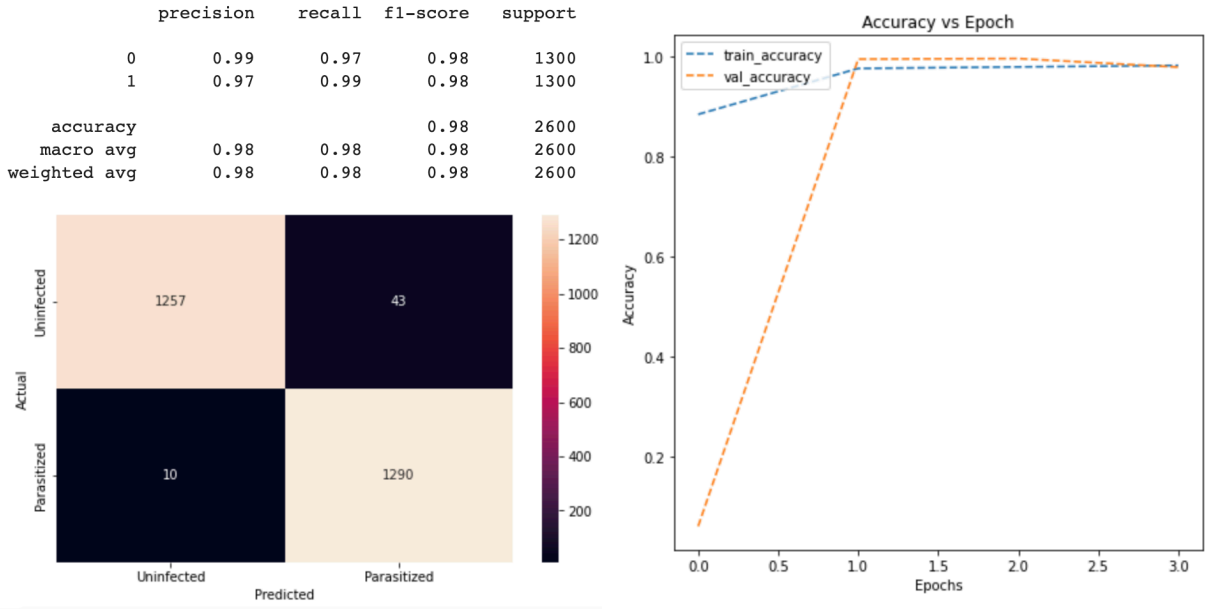
Malaria is a life-threatening illness, usually found in tropical countries, that is caused by Plasmodium parasites. The parasites come from infected female mosquitoes and can spread through to people via bites of the infected mosquito[1]. Symptoms of malaria include fever, headache and chills and the severe malaria stages include multi-organ failure and respiratory distress or anaemia. In 2021, nearly half the world's population was at risk of malaria[2]. Malaria is a preventable and treatable illness, however, it requires prompt diagnosis to administer the treatment. Therefore, early and accurate prevention of malaria can help drastically reduce the deaths caused by this disease.

Malaria is usually diagnosed through microscopic blood smears or rapid diagnostic tests. For the microscopic diagnosis, a blood smear is taken on the microscope slide and is stained (often with a Giemsa stain) to allow the Plasmodium parasite to show up clearly in the cell's image[3]. The goal of this project is **to develop a fast and accurate computer vision model to detect malaria using a dataset of images of the stained blood smears**. The neural network model presented here is able to accurately classify cellular images as "Parasitized" or "Uninfected" depending on if the cells contain the Plasmodium parasite or not, respectively. Using that trained model, the model was able to predict whether cells were Parasitized or Uninfected which can help in the early and quick detection of Malaria.

Solution design

The proposed solution would be to implement Model 2 as a method of solving the problem of diagnosing Malaria quickly and accurately. Model 2 is a Convolutional Neural Network (CNN) that was trained on labeled images of Parasitized and Uninfected cells. As a CNN, it is able to

pick up minute differences in the image in order to correctly to label the cells as Parasitized or Uninfected to an accuracy of 98.77%, which makes it a feasible and accurate model to use to solve the problem (***Fig. 1***)



	precision	recall	f1-score	support
0	0.99	0.97	0.98	1300
1	0.97	0.99	0.98	1300
accuracy			0.98	2600
macro avg	0.98	0.98	0.98	2600
weighted avg	0.98	0.98	0.98	2600

a b

Figure 1 Model 2 a) Classification Report and Confusion Matrix and b) Model Accuracy

Model 2 consists of 21 layers that include the Convolution, Leaky ReLU, MaxPooling, Batch Normalization, Dense and a Flatten and Dropout Layer. Only one Dropout layer was used to prevent the validation accuracy from being higher than the training accuracy (***Fig. 1b***). The multiple Convolutional Layers allow the model to learn the spatial component and translational invariance of the images, as it was observed that the parasite in the parasitized cells might be in different locations of the image. The Leaky ReLU layers allows neurons to be activated, unlike the Relu, which doesn't allow values under 0 and therefore doesn't block the learning. Batch normalization was chosen to normalize the contributions to a layer for every mini-batch, making learning quicker and decreases the number of epochs required. Max Pooling allowed the important features of the image to be retained, which is helpful for a detailed and specific image such as the ones in this dataset. Overall, due to the size of the dataset and the deep CNN, the model was able to train on the data properly and learn the specifics of the features.

Model 2 was also trained with an Adam optimizer with a learning rate of 0.001. Callbacks were implemented in this model which stopped training once validation loss stopped improving. A batch size of 32 and 20 epochs were used in training. The validation split was 0.25, which meant that 25% of the total data used in training was split into the validation dataset.

Many of the models trained (illustrated in ***Appendix 2***) are around the same range for the test accuracy. However, Model 2 had the highest a test accuracy of 0.9877 which makes it the most optimal choice for the solution to the problem statement (***Fig. 1***).

Comparisons to other trained models also resulted in Model 2 being chosen as the best choice solution. The Base Model has a similar test accuracy to Model 2, however, the recall for the parasitized cells is 0.99 (Model 2) vs 0.98 (for Base Model) ***(Fig. 1, Appendix 3)***. In comparison,

Model 3, which was the model trained on the HSV images of the dataset, also had a very similar test accuracy to Model 2, with better recall Label “0”, uninfected images. The Base Model predicted 24 parasitized cells as uninfected, in comparison to 15 for Model 1, 10 for Model 2 and 18 for Model 3 ***(Appendix 3-5)*** . The Image Augmentation Model, trained on an augmented dataset incorporating translations and rotations, and the Transfer Learning model, trained on a VGG16 model with pre-built architecture, predicted 61 vs. 65 parasitized cells as uninfected from the test dates, respectively (***Appendix 6-7***). The Image Augmentation Model and Transfer Learning models also have the lowest accuracy and recall rates for the labels for parasitized cells, showing that their performance is poor in comparison to the other four models.

Analysis and Key Insights

Model 2, as a CNN, takes in images of the dyed cells from a blood smear and can output a prediction on whether the cells are uninfected or parasitized. However, as the neural network was trained on specific images of the cells from the dataset, the images fed into the network for a prediction output will need to be images taken the same way with the same dye as the training images (***Fig. 2***). This makes the model a bit less robust as image qualities such as contrast, resolution and more will heavily impact the outcome of the model.

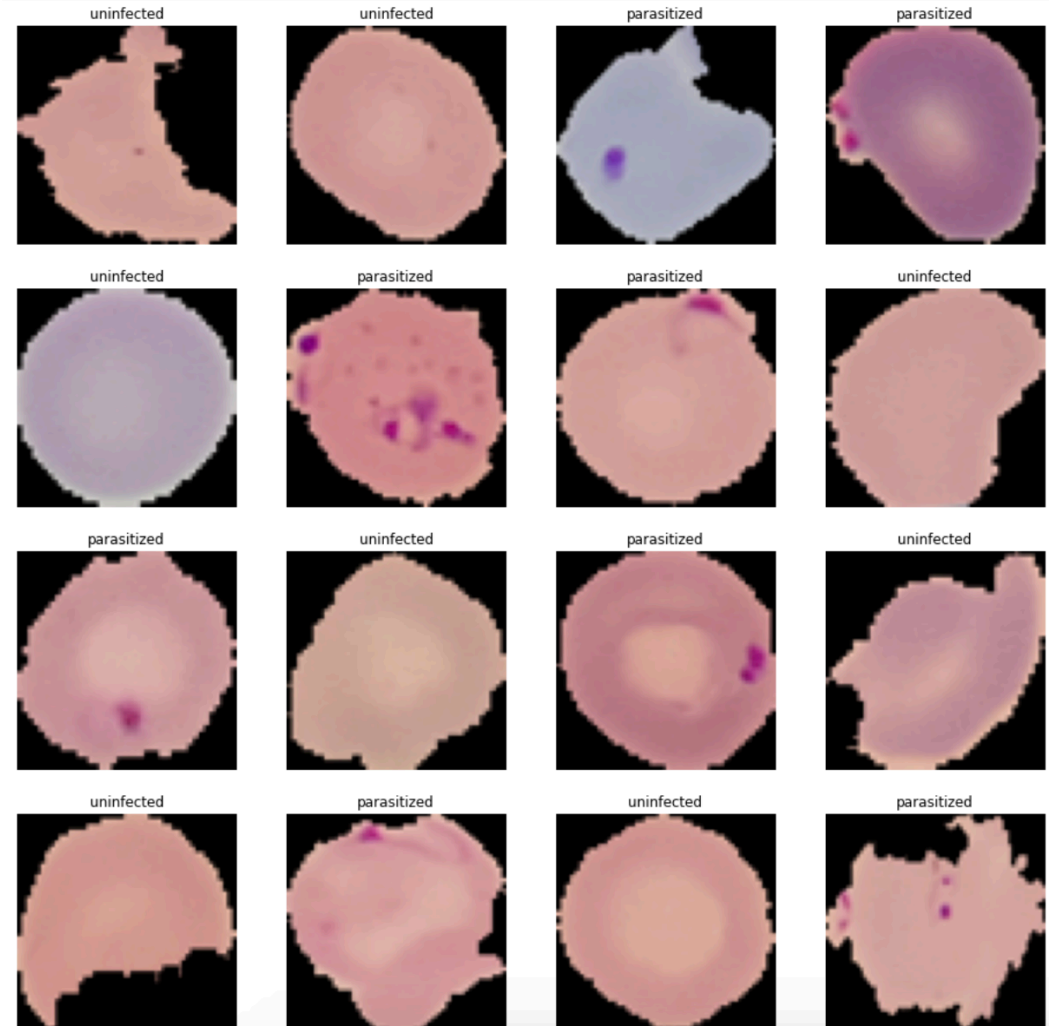


Figure 2 Visualization of a selection of images from the training dataset along with their labels

However if similar images as training images are used, then Model 2's high recall for parasitized cells of 0.99 means that cells that are parasitized will rarely be misclassified as uninfected, which limits false negative diagnoses. Additionally, it's overall accuracy will also help prevent false positives (also seen by the recall of 0.97 for uninfected cells).

As Model 2 is an automated way of detecting if the Malaria parasite exists in the cells of the sample, it would make the process of diagnosing Malaria a lot quicker. The patient would not have to wait for a pathologist to analyze the histology images and would instead get an accurate answer on whether the Plasmodium parasite exists within their cells, showing that the patient has Malaria. This would be a very beneficial solution to providing a quick and early diagnosis to a patient possibly suffering from Malaria. Due to this, patients who have Malaria can get treated earlier and have a higher chance of surviving the illness.

Recommendations for implementation

In order to implement Model 2 as a solution to the problem of diagnosing Malaria, it is recommended and necessary to have stained images of the blood smears be provided so that the algorithm can clearly detect if the parasite is present in the cells. Stakeholders will need to get possible patients to a histology lab so that these blood smears can be taken and the images will need to be sent over online to the computer where this Model is deployed. Afterwards, the model can be run on the dataset of images provided by the histology lab so that a quick but accurate answer on whether the patient has Malaria can be provided. The benefits of using this model would be that the patient won't have to wait for the results analyzed by a pathologist to identify if they have Malaria. The algorithm will provided a quicker method to determine if they have the Malaria parasite or not. In case the algorithm states that the patient does have Malaria, then treatment can be started earlier which can help mitigate the symptoms of the illness. If needed, the images can also be sent to a pathologist in the meanwhile to confirm if the algorithm was accurate, that way any possible false negatives missed by the algorithm may also be caught.

The cost of this algorithm is that it does not have the knowledge of a trained pathologist. It is using images and pattern recognition to determine if the cells contain the parasite, therefore, it has the possibility of having false negative results where it can misclassify a Parasitized cell as an Uninfected cell. In these cases, the patient will not be treated early for Malaria which can cause more severe symptoms. However, in the case of Model 2, the recall is at 99%, which means that it will not result in as many false negative as the other models. The Model can misclassify 1% of the time, which is the innate risk of using this machine learning model for diagnosing Malaria. However, the benefit of using the model still outweighs the risk as the Model is performing at 98.77% accuracy.

In order to use this model, the same dye that is used in the training images will need to be used for the dataset requiring prediction. The images of the blood smear will also need to be taken in the same way as the training dataset images. In case of any large image differences, the model will not perform the best it can. Therefore, the challenges of this approach would be to ensure that all the histology images are taken the same way as the images in the training dataset. The

correct magnification and image size will also need to be taken into account when using Model 2 to predict the classification of Parasitized or Uninfected. Overall, the images will need to have the same image qualities as the training dataset which includes image qualities such as contrast, brightness, saturation and resolution.

As it may be hard to have the images with the exact same qualities as the training images, an improvement could be to develop an algorithm that is more applicable to different image qualities. Further analysis that can be done is exploring image segmentation as a step to use in building a model. If images with a different contrasts or magnification etc. are taken for the blood smears, a model built with image segmentation can help segment out the parasite in the Parasitized cells and then those images can be used to train a model which can in turn be used to predict on different images that aren't necessarily taken the same way (after applying segmentation to the test images). Using Model 2 would be a good starting point to diagnose Malaria early on without having to wait for a pathologist but the images will need to be taken the same way as images in the training dataset, which may create a challenge. However, for a disease as prevalent as Malaria, providing a quick solution to diagnosing the disease will allow earlier treatment which will make a strong impact on the survival rates.

Bibliography

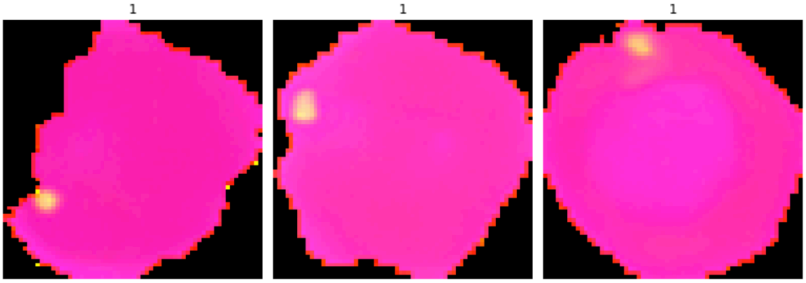
1. World Health Organization. (n.d.). Malaria. World Health Organization. Retrieved March 27, 2023, from https://www.who.int/news-room/questions-and-answers/item/malaria?gclid=CjwKCAjw_YShBhAiEiwAMomsEAJbrNJ1dZEzL-f7JNH6GtSakdN_6-3TNF-Fe5SIIMUanv5wGWqZ0xoCqpMQAvD_BwE

2. World Health Organization. (n.d.). Fact sheet about malaria. World Health Organization. Retrieved March 27, 2023, from https://www.who.int/news-room/fact-sheets/detail/malaria#:~:text=Key%20facts,million%20cases%20of%20malaria%20worldwide.

3. Centers for Disease Control and Prevention. (2018, July 23). CDC - Malaria - diagnosis & treatment (United States) - diagnosis (U.S.). Centers for Disease Control and Prevention. Retrieved March 27, 2023, from https://www.cdc.gov/malaria/diagnosis_treatment/diagnosis.html#:~:text=Malaria%20parasites%20can%20be%20identified,the%20parasites%20a%20distinctive%20appearance.

Appendix

Appendix 1: HSV Images



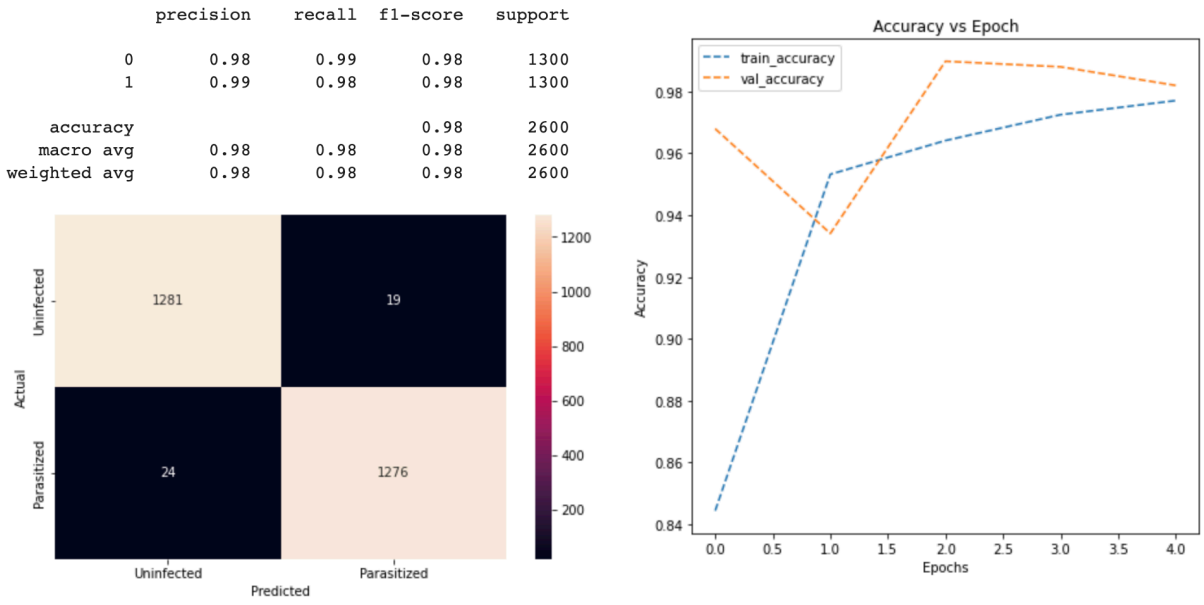
Appendix 1 Examples of images converted to the HSV scale

Appendix 2: Comparison of Model Performance

Model	Recall Label "0" Uninfected Cells	Recall for Label "1" Parasitized Cells	Test Accuracy
Base Model	0.99	0.98	0.9835
Model 1	0.96	0.99	0.9758
Model 2	0.97	0.99	0.9877
Model 3 - HSV Trained	0.98	0.99	0.9838
Image Augmentation	0.99	0.95	0.9723
Transfer Learning	0.95	0.95	0.9512

Appendix 2 Comparison of Metrics for the Different Models

Appendix 3: Base Model Classification Report, Confusion Matrix and Accuracy

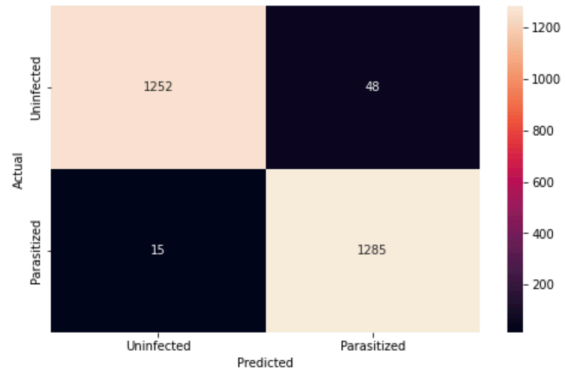


a b

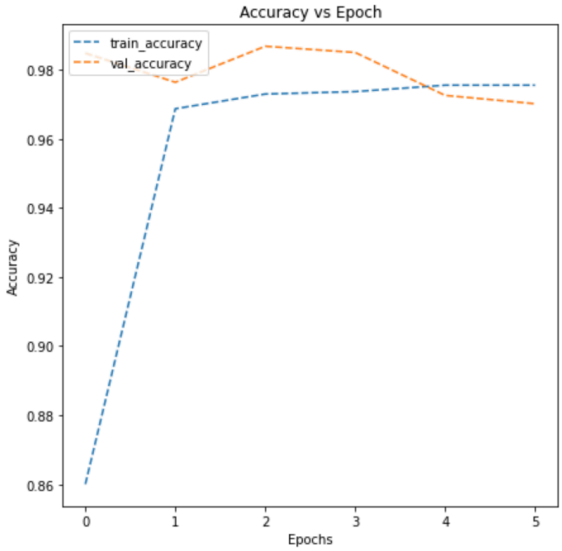
Appendix 3 Base Model a) Classification Report and Confusion Matrix and b) Model Accuracy

Appendix 4: Model 1 Classification Report, Confusion Matrix and Accuracy

	precision	recall	f1-score	support
0	0.99	0.96	0.98	1300
1	0.96	0.99	0.98	1300
accuracy			0.98	2600
macro avg	0.98	0.98	0.98	2600
weighted avg	0.98	0.98	0.98	2600



a

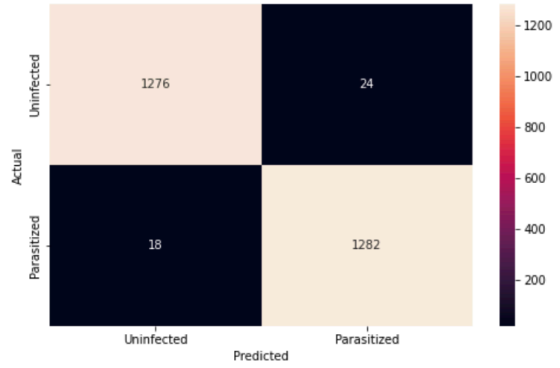


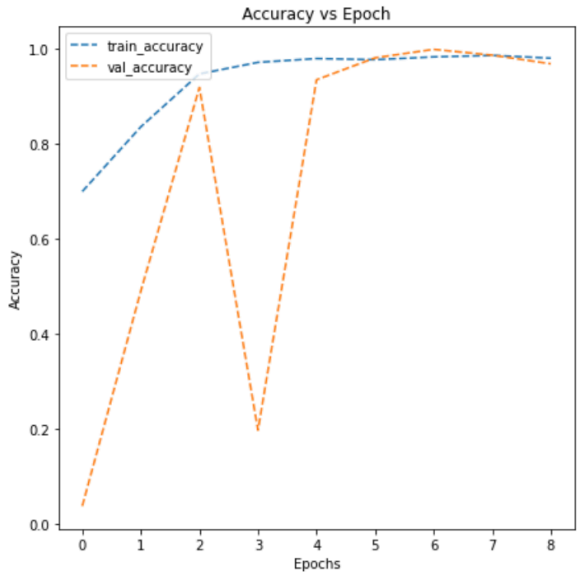
b

Appendix 4 Model 1 a) Classification Report and Confusion Matrix and b) Model Accuracy

Appendix 5: Model 3 (HSV Images Trained) Classification Report, Confusion Matrix and Accuracy

	precision	recall	f1-score	support
0	0.99	0.98	0.98	1300
1	0.98	0.99	0.98	1300
accuracy			0.98	2600
macro avg	0.98	0.98	0.98	2600
weighted avg	0.98	0.98	0.98	2600

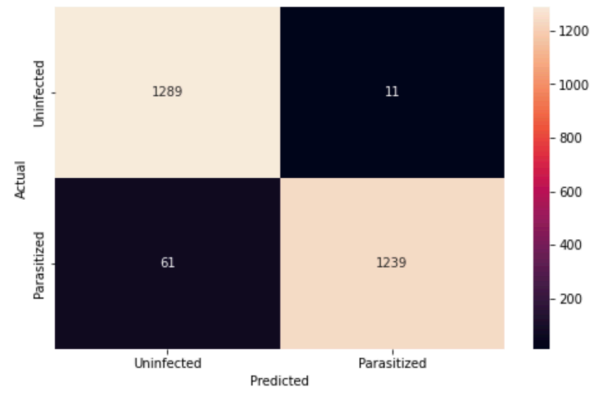




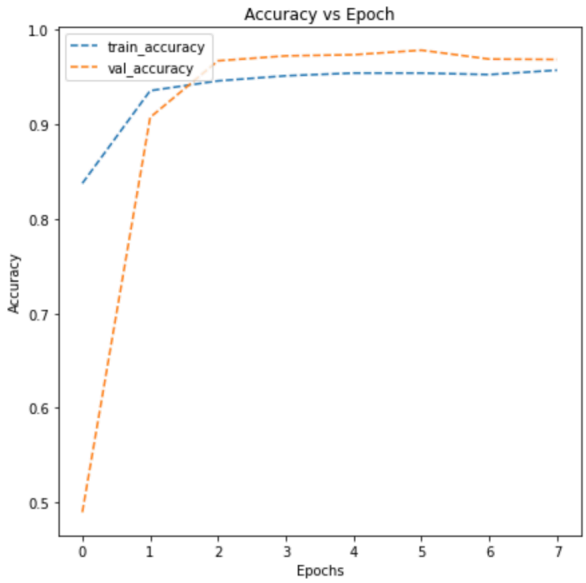
Appendix 5 Model 3 a) Classification Report and Confusion Matrix and b) Model Accuracy

Appendix 6: Image Augmentation Model Classification Report, Confusion Matrix and Accuracy

	precision	recall	f1-score	support
0	0.95	0.99	0.97	1300
1	0.99	0.95	0.97	1300
accuracy			0.97	2600
macro avg	0.97	0.97	0.97	2600
weighted avg	0.97	0.97	0.97	2600



a

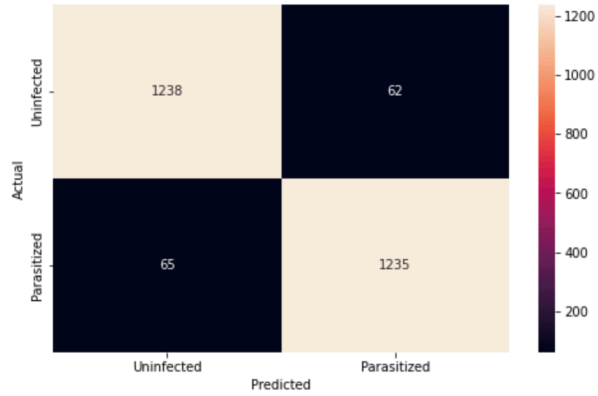


b

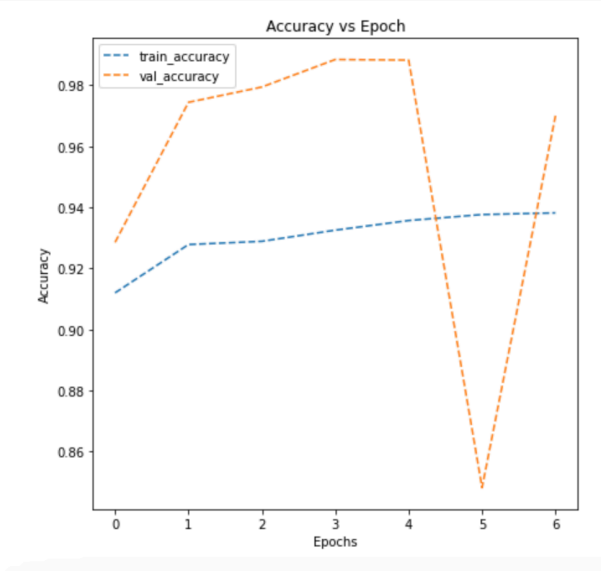
Appendix 6 Image Augmentation Model a) Classification Report and Confusion Matrix and b) Model Accuracy

Appendix 7: Transfer Learning Model Classification Report, Confusion Matrix and Accuracy

	precision	recall	f1-score	support
0	0.95	0.95	0.95	1300
1	0.95	0.95	0.95	1300
accuracy			0.95	2600
macro avg	0.95	0.95	0.95	2600
weighted avg	0.95	0.95	0.95	2600



a



b

Appendix 7 Transfer Learning a) Classification Report and Confusion Matrix and b) Model Accuracy