Capstone Project

Malaria detection

Olatz Terreros
April 14, 2023

Capstone Project: Malaria detection

1. Executive summary

This project proposes a model based on convolutional neural networks (CNNs) for the detection of malaria. The model identifies whether the image of a red blood cell corresponds to an infected one or not, and classifies it as parasitized or uninfected. The model provides an overall accuracy of 98.8% for the analyzed data.

The model could improve the accuracy and speed of diagnosis, which is critical for effective treatment and disease management. Healthcare professionals could make faster and more informed decisions about patient care, which could help to reduce the spread of malaria and improve outcomes for those affected by it.

The false negative and false positive rates of the model should be considered. False negatives can result in missed diagnoses, while false positives can lead to unnecessary treatment and increased costs. Both rates should be minimized as much as possible. It is therefore recommended to increase the size and diversity of the dataset, as well as classify images according to malaria development stages.

2. Problem summary

Malaria is a disease caused by a parasite called Plasmodium, which is transmitted to humans through the bites of infected female Anopheles mosquitoes. Symptoms of malaria typically include fever, chills, headache, muscle aches, and fatigue. In severe cases, the disease can lead to complications such as anemia, kidney failure, seizures, and coma, and it can even be fatal.

The disease is prevalent in many tropical and subtropical regions of the world, including sub-Saharan Africa, Southeast Asia, and Latin America. Almost 50% of the world's population is in danger from malaria. It is a significant global health problem, with over 247 million cases and nearly 619.000 deaths reported in 2021, primarily in sub-Saharan Africa [1]. Children under 5 years of age are the most vulnerable population group affected by malaria; in 2018 they accounted for 67% of all malaria deaths worldwide [2].

Malaria affects not only the health of individuals but also the economic development of countries. The disease can lead to decreased productivity and increased healthcare costs.

Traditional diagnosis of malaria in the laboratory requires careful inspection by an experienced professional to discriminate between healthy and infected red blood cells. It is a tedious, time-consuming process, and the diagnostic accuracy (which heavily depends on human expertise) can be adversely impacted by inter-observer variability.
The model presented in this project has the potential to improve the accuracy, efficiency, and cost-effectiveness of malaria diagnosis, preventing the progression of the disease and improving patient outcomes.

3. Solution design

Several models were explored as part of the solution design (table 1).

Table 1: Model settings and metrics

Model	Normalized Prediction Time	Accuracy (%)	Recall Uninfected (%)	Recall Parasitized (%)	False negatives	False positives	True Negatives	True Positives	Description
1	0,21	97.6	97.8	97.3	35	28	1272	1265	• 3 conv. layers • No batch normalization • Act. function ReLU
2	0,16	98.3	98	98.7	17	26	1274	1283	• 5 conv. layers • No batch normalization • Act. function ReLU
3	0,19	98.6	99	98.2	23	13	1287	1277	• 5 conv. layers • Batch normalization • Act. function LeakyReLU
4	1,00	98.8	98.7	98.8	15	17	1283	1285	• Model 3 architecture • Data Augmentation (horizontal flip, vertical flip, rotation: 35°)
5	0,37	97.2	97	97.5	33	39	1261	1267	• Model 3 architecture • Data Augmentation (RGB to HSV)
6	0,28	98.5	98.4	98.5	19	21	1279	1281	• Model 3 architecture • Data Augmentation (Gaussian Blurring)
7	1,00	95	97.8	92	104	29	1271	1196	• Transfer learning (VGG16)

According to the results presented in table 1, model 4 should be adopted for malaria detection. It provides the highest accuracy (98.8%) and the lowest number of false negatives (15). Reducing false negatives in malaria detection with CNN (Convolutional Neural Networks) is important since false negatives can result in incorrect diagnosis and

treatment of patients, which can have serious consequences, including the possibility of death.

The following data augmentation techniques were applied to the dataset:

- Random rotations: rotating the image randomly can help to increase the robustness of the model to variations in the orientation of the blood cell. Rotation range used: 35 °.
- Horizontal and vertical flip: flipping the image horizontally and vertically can help to increase the diversity of the dataset and improve the performance of the model.

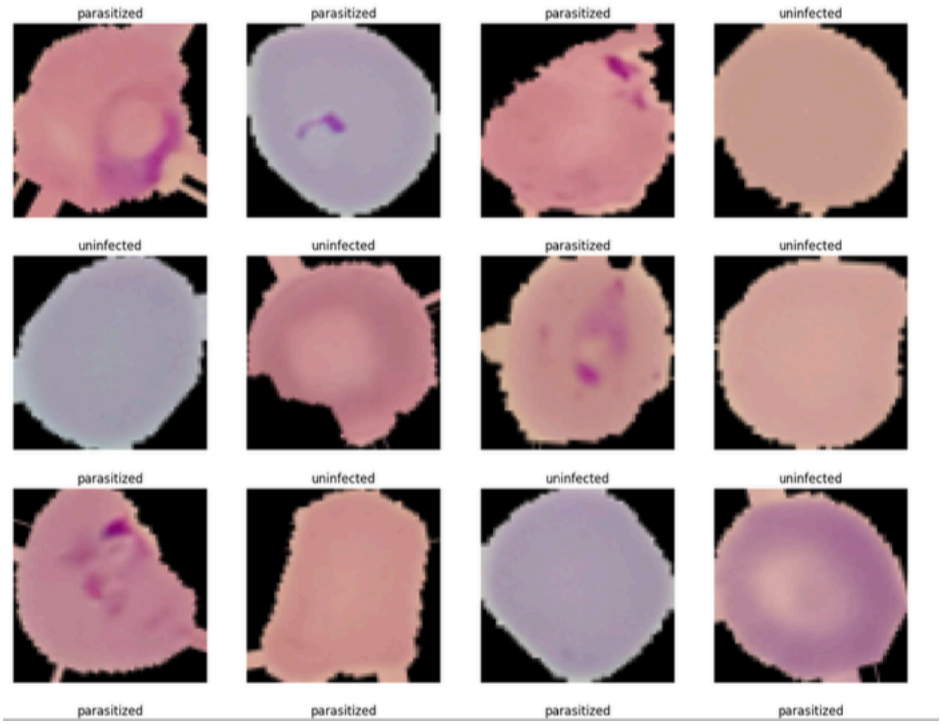

Figure 1: Sample of red blood cells (parasitized and uninfected) after implementing data augmentation (rotation 35°, horizontal and vertical flipping).

Model 4 is a CNN model with 5 convolutional layers followed by two fully connected (dense) layers and an output layer.

The first convolutional layer has 32 filters with a kernel size of (3,3) and uses a LeakyReLU activation function with an alpha value of 0.1. It is followed by batch normalization and a max pooling layer with a pool size of (2,2). The second convolutional layer has 64 filters with the same kernel size, activation function, normalization, and max pooling as the first layer.

The third convolutional layer has 128 filters with the same kernel size, activation function, normalization, and max pooling as the first two layers. The fourth

convolutional layer has 256 filters with the same kernel size, activation function, normalization, and max pooling as the previous three layers. The fifth and final convolutional layer has 512 filters with the same kernel size, activation function, normalization, and max pooling as the previous layers.

After the final convolutional layer, the output is flattened and fed into two fully connected layers with 512 and 256 nodes, respectively. Each dense layer uses a LeakyReLU activation function, batch normalization, and dropout regularization with a rate of 0.5. The output layer has 2 nodes with a softmax activation function, indicating the probability of each input belonging to one of two classes.

The model is compiled using binary cross-entropy as the loss function, Adam optimizer with a learning rate of 0.001, and accuracy as the evaluation metric.

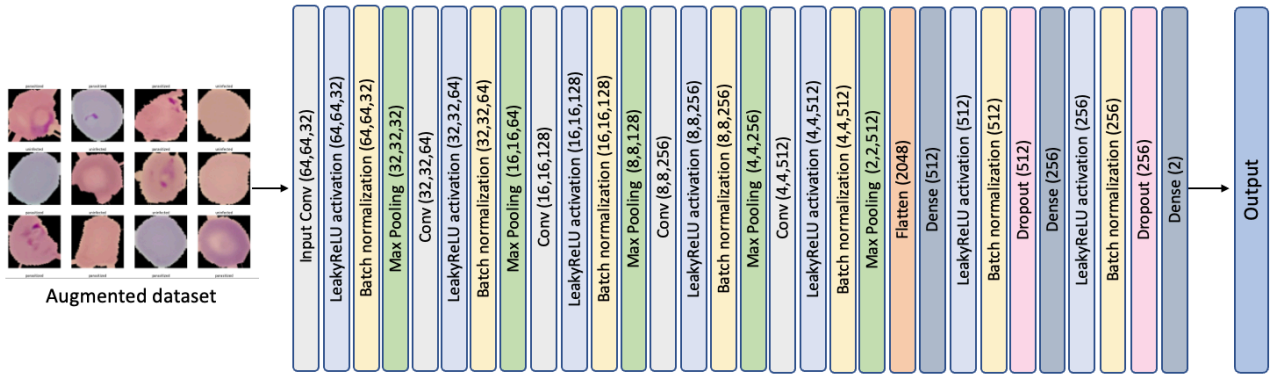

Figure 2: Architecture of model 4

4. Recommendations for Implementation

The proposed model should be deployed in the laboratory by integrating it into the laboratory information system (LIS) or a custom software application. The model should be accessible to laboratory technicians who can upload blood cell images and receive the corresponding diagnosis. A quality assurance and control program should be implemented to ensure that the model is working as expected and that the diagnoses are accurate. Training and support should be provided to laboratory technicians on how to use the CNN model and interpret the results. Training materials and documentation should be developed to help new users use the model.

The integration of the CNN model into a laboratory for malaria diagnosis presents several benefits.

- Accuracy: the CNN model can be highly accurate in detecting malaria parasites in blood cell images. It can be trained on large datasets of labeled images and can learn to recognize patterns that are indicative of infected cells.
- Speed: the model can process large volumes of images quickly and efficiently, allowing for rapid diagnosis of malaria infections. This can reduce the time required for laboratory technicians to analyze blood cell images manually, allowing for more efficient use of resources.
- Consistency: the model can provide consistent diagnoses, regardless of the experience or skill level of the laboratory technician. This can reduce the likelihood of errors and improve the reliability of malaria diagnoses.
- Cost-effectiveness: the model can be run on inexpensive hardware, making it a cost-effective solution for malaria diagnosis. It can also reduce the need for expensive laboratory equipment and supplies.
- Flexibility: the model can be adapted to work with a wide range of imaging equipment and settings, making it a flexible solution for malaria diagnosis. It can also be trained to detect other types of blood-borne pathogens, making it useful for a variety of laboratory applications.

While the proposed model has shown promising results, there are still some challenges that need to be addressed:

- Availability of high-quality data: CNN models require large amounts of high-quality data to be trained effectively. However, obtaining such data for malaria diagnosis can be challenging, especially in resource-constrained settings where the disease burden is highest.
- Generalization to new settings: CNN models trained on data from one region may not perform as well in other regions due to differences in the prevalence of different malaria species, variations in image quality, and other factors.
- Interpretability: CNN models are often referred to as "black boxes" because it can be difficult to understand how they arrive at their predictions. This lack of interpretability can be a challenge for healthcare providers who need to make treatment decisions based on the model's output.
- Ethical considerations: The use of AI technology for healthcare diagnosis raises ethical concerns, such as privacy, security, and bias in the model's output. These concerns need to be addressed to ensure that the technology is used in a responsible and ethical manner.
- Costs of deployment and maintenance: Once the CNN model is developed, it needs to be deployed and maintained. This can involve integrating the model with other diagnostic methods and developing a user-friendly interface. The cost of deployment and maintenance can depend on the complexity of the system and the ongoing maintenance requirements.

Improving the proposed model for malaria diagnosis will require ongoing research and development efforts, as well as collaboration between researchers, healthcare providers, and other stakeholders. The following aspects should be considered:

- Reduction of false negatives: the misclassified images should be examined to identify common patterns or features that may have caused the model to make an incorrect prediction. The model or the data should be adapted according to the outcome of the analysis.

  Parasitized cells can exhibit distinct differences in both color and shape when compared to uninfected ones, depending on the life cycle stage of the malaria parasite (figure 3).

  In the early stages of malaria infection, parasitized red blood cells may appear normal and have a similar color to uninfected red blood cells. However, as the parasite multiplies within the red blood cells, the cells can become pale and lose their typical red color. This is due to the hemoglobin, which is the protein responsible for carrying oxygen in red blood cells, being broken down by the parasite. As a result, infected red blood cells can appear more transparent or "ghost-like" than uninfected cells.

  The malaria parasite can also cause changes in the shape of red blood cells. In the early stages of infection, the infected red blood cells may appear slightly larger and more rounded than uninfected cells. As the infection progresses, the infected red blood cells can become more irregular in shape and can form a variety of distorted shapes, such as crescent shapes, oval shapes, or multiple lobes. This is due to the parasite modifying the red blood cell membrane to facilitate its own survival and replication within the cell.

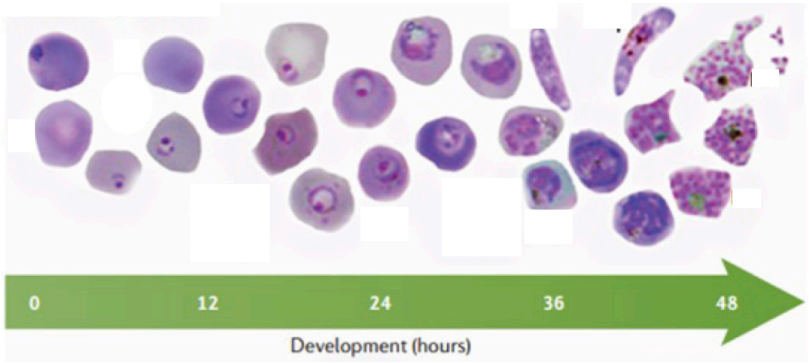

Figure 3: Malaria parasite life cycle

The false negatives occur mostly with either the first or last stage of parasite life cycle. False positives look infected at first glance (figure 4).

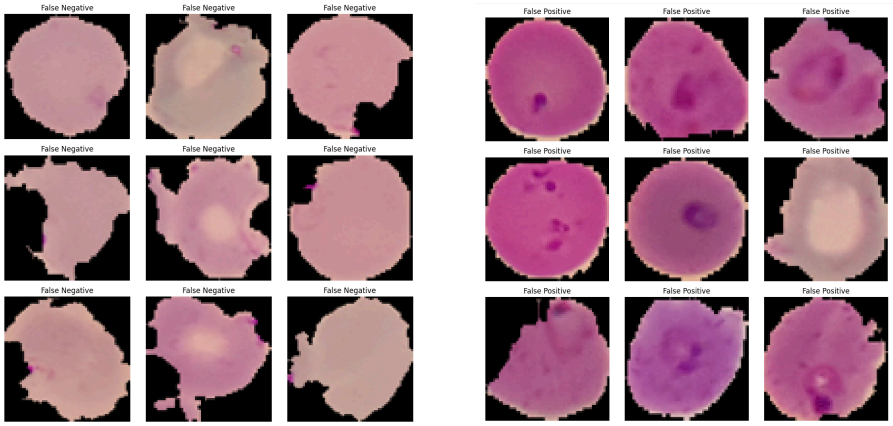

Figure 4: False negatives (left) and false positives (right)

The images of our dataset can be grouped based on the life cycle of the malaria parasite inside the cell. This could reduce the level of misclassification of our model.

- Use of transfer learning: the model proposed requires a higher computational time than the other models analyzed. Transfer learning involves using a pre-trained CNN model and fine-tuning it on a new dataset. This approach can save time and resources compared to training a CNN model from scratch and can help improve the model's accuracy. The model VGG16 does not give good results, probably because it is trained on images that do not resemble those in our database. Another pre-trained model could be investigated.

- Increase the dataset: a CNN model trained on a diverse range of data is more likely to perform well on new data. Collecting data from different countries and regions can help increase the diversity of the training data, improving the model's accuracy. Different species of malaria can have different morphological characteristics, which can impact the accuracy of a CNN model. Therefore collecting data from different types of malaria can help ensure that the model is able to accurately detect a wide range of malaria types.

5. Bibliography

[1] «World Health Organization,» 2023. [En línea]. Available: https://www.who.int/news-room/fact-sheets/detail/malaria.

[2] «UNICEF,» 2023. [En línea]. Available: https://www.unicef.org/health/childhood-diseases.