Malaria Detection via SAM-based Red Blood Cell Segmentation and Feature Analysis

Anna Mur Suñé

Master's Degree in Bioinformatics and Biostatistics

Statistical Bioinformatics and Machine Learning

Advisor

Edwin Santiago Alférez Baquero

June 2024

FITXA DEL TREBALL FINAL

Title:	*Malaria Detection via SAM-based Red Blood Cell Segmentation and Feature Analysis*
Author:	*Anna Mur Suñé*
Advisor:	*Edwin Santiago Alférez Baquero*
SRP:	*Edwin Santiago Alférez Baquero*
Date of delivery (mm/aaaa):	*06/2024*
Studies:	*Master's Degree in Bioinformatics and Biostatistics*
Area:	*Statistical Bioinformatics and Machine Learning*
Language:	*English*
Keywords	*Malaria detection, machine learning, red blood cell segmentation, SAM, feature importance*

Abstract

The diagnosis of malaria is still predominantly based on the manual observation of blood samples through optical microscopy, a time-consuming and tiring task for laboratory staff. As an alternative to this gold standard, previous studies suggest that the application of artificial intelligence can allow for faster and more accurate diagnoses, leading to more effective patient treatment.

This study aims to develop a pipeline for the automatic image-based detection of malaria using a traditional machine learning method to classify Plasmodium-infected and uninfected red blood cells. Additionally, the project seeks to identify the most important types of image features that enable differentiation between infected and uninfected cells.

The methodology includes the manual annotation of red blood cells from thin blood smear images of malaria patients, segmentation of cells using the novel Segment Anything Model (SAM), application of the Random Forest algorithm to classify segmented cells into infected or uninfected classes using a feature dataset, and calculation of feature importances. Moreover, an object detection model is trained to automate the detection of RBCs in microscopy images.

Results of segmentation with SAM show remarkable performance. After training and testing the classification model within an eighty-feature dataset of colour, texture, and morphology, the accuracy and F1-score obtained are 99.5% and 99.2%, respectively. Thus, an accurate malaria diagnosis could be achieved by applying this pipeline based on SAM segmentation and Random Forest classification to analyse thin smears. Regarding feature importance, colour features, specifically green and red histograms, appear to be the most distinctive.

Contents

1 INTRODUCTION 1
1.1 CONTEXT AND JUSTIFICATION 1
1.2 GENERAL DESCRIPTION 1
1.3 IMPACT ON SUSTAINABILITY, ETHICAL-SOCIAL AND DIVERSITY 2
1.3.1 Sustainability 2
1.3.2 Ethical behaviour and social responsibility 2
1.3.3 Diversity and human rights 2
1.4 OBJECTIVES 3
1.4.1 General objectives 3
1.4.2 Specific objectives 3
1.5 APPROACHES AND METHODOLOGY 3
1.6 WORK PLANNING 5
1.6.1 Tasks and milestones 5
1.6.2 Calendar 6
1.6.3 Risk analysis 6
1.7 BRIEF SUMMARY OF THE OBTAINED PRODUCTS 6
1.8 BRIEF DESCRIPTION OF THE OTHER CHAPTERS OF THE REPORT 7
2 STATE OF THE ART: MALARIA DIAGNOSIS 8
2.1 TRADITIONAL METHODS 8
2.2 ARTIFICIAL INTELLIGENCE METHODS 9
2.2.1 Machine learning 9
2.2.2 Deep learning 10
3 MATERIALS AND METHODS 11
3.1 TOOLS 11
3.2 THE DATASET 11
3.3 ANNOTATION 11
3.4 SEGMENTATION 12
3.5 FEATURE EXTRACTION 13
3.5.1 Colour features 14
3.5.2 Texture features 14
3.5.3 Geometric features 15
3.6 MACHINE LEARNING CLASSIFICATION 15
3.6.1 Random forest algorithm 15
3.7 FEATURES EVALUATION 17
3.7.1 Feature importance: Gini 17
3.7.2 SHAP values 18

3.8 OBJECT DETECTION MODEL 18
4 RESULTS AND DISCUSSION 19
4.1 ANNOTATION 19
4.2 SEGMENTATION 20
4.3 FEATURES EXTRACTION 21
4.4 MACHINE LEARNING CLASSIFICATION 22
4.5 FEATURES EVALUATION 24
4.5.1 Features importance 24
4.5.2 SHAP values 25
4.5.3 Interpretation of feature importance 27
4.6 OBJECT DETECTION MODEL 28
5 CONCLUSIONS 29
5.1 FUTURE WORK 29
6 GLOSSARY 31
7 BIBLIOGRAPHY 32
8 APPENDIX 35
8.1 SEGMENTATION EFFICIENCY 35
8.2 BINARY IMAGES OPTIMIZATION 35
8.3 MISSCLASSIFIED IMAGES IN THE CLASSIFICATION MODEL 36

List of figures

Figure 1. Sustainable Development Goals for 2030. 2
Figure 2. Main steps followed for automated malaria diagnosis. 3
Figure 3. Gantt chart of the project planning. 6
Figure 4. A) Preparations of thick and thin blood smears (17). (B) Microscopic images of thick and (C) thin smears. (18) 8
Figure 5. Life cycle of malaria (19). 8
Figure 6. Scheme of artificial intelligence organization. 10
Figure 7. Image annotation in Label Studio 12
Figure 8. Segment Anything Model (SAM) overview (5). 13
Figure 9. Structure of dataset, which includes features, class and image path for each image. 15
Figure 10. Random forest scheme (31). 16
Figure 11. Confusion matrix. 17
Figure 12. Example of the different type of cells found in the samples. 19
Figure 13. Annotated image with LabelStudio visualized in Google Colab. 19
Figure 14. Comparison of segmented images from patient 5 with SAM (A) without multimask option or (B) with multimask option. 20
Figure 15. Image overlayed with the three masks obtained with the multimask option of SAM. 20
Figure 16. Visually verification of segmentation. A) source image with bounding boxes, B) definitive mask (the one with a highest score) and C) image overlapping. 20
Figure 17. Overlap of source image and mask (A) and final image (B). 20
Figure 18. Example of segmented RBCs. 21
Figure 19. Example of not entire uninfected RBCs. 21
Figure 20. Example of A) infected and B) uninfected RBC images in RGB, red, green and blue components from RGB image, grayscale image and binary image. 21
Figure 21. Example of A) infected and B) uninfected RBC image in RGB and their corresponding histograms in red, green and blue components. 22
Figure 22. Features importance bar plot. 24
Figure 23. Summary plot of SHAP 25
Figure 24. Global SHAP summary plots for A) uninfected class and B) infected class. 26
Figure 25. Local SHAP of 5 instances from uninfected class. 26
Figure 26. Local SHAP of 5 instances from infected class. 27
Figure 27. Evaluation images of the object detection model. 28

List of tables

Table 1. Start and end dates of the five derivable tasks. 5
Table 2. List of tasks and subtasks, their milestones, and duration in days. 5
Table 3. Summary of the colour, texture, and geometric features extracted 13
Table 4. Comparison of data before and after undersampling. 23
Table 5. Model performance metrics 23
Table 6. Features importance values of the 20 most important features. 24

1 Introduction

1.1 Context and justification

Malaria is one of the most life-threatening diseases caused by Plasmodium parasites and transmitted by infected female Anopheles mosquitoes. According to the World Health Organization (WHO), millions of people are infected annually, and it is responsible for thousands of deaths in the infected population, nearly all in sub-Saharan Africa. In 2022, 608.000 deaths and 249 million cases of malaria were reported globally, an increase of 5 million cases compared to 2021 (1). Investment in vector control tools, vaccines, antimalarial drugs, and new diagnostics is required to accelerate progress against malaria (2).

The gold standard technique for detecting malaria is optical microscopy, which requires the manual checking of patient samples. Not only the identification of the parasite but also the counting of parasites is necessary to know the severity of the disease and the efficacy of drug treatment over time. However, the main drawback of the technique is that it needs a highly trained personal, and it is a time-consuming task, as around 15-30 minutes are required by an expert to examine only one blood smear (3).

As a breakthrough, artificial intelligence (AI) has become as an essential tool in diagnosis from medical images that can also be applied as an alternative technique to obtain a more reliable diagnosis for malaria in a real work environment, where human error in traditional microscopy has been identified as a major cause of misdiagnosis (4). Additionally, having a faster diagnosis provides timely treatment, allowing these new tools to have direct consequences in achieving safe patient care, reducing the risk of complications, and avoiding death.

The main aim of this work is to develop a pipeline that allows an automatic, faster, and more accurate malaria detection from patient peripheral blood smear images as well as identify which are the most important features of the images that enable to confirm this disease.

1.2 General description

The project involves developing a pipeline for malaria detection using blood smear images. The proposed method includes the following steps: image acquisition, image annotation, red blood cell (RBC) segmentation, feature extraction, classification, and analysis of feature importance.

Specifically, the novel Segment Anything Model (SAM) (5) was applied for RBCs segmentation. Colour, texture and geometric image feature extraction was then performed. Random forest algorithm was employed for image classification into infected or uninfected classes. The last step was an interpretive analysis of feature importance to determine which characteristics are crucial for malaria detection. Additionally, an object detection model was trained to locate RBCs within microscope images and complete the computer vision pipeline.

The report work includes the current state-of-the-art in the field of malaria diagnosis, as well as the methodology, results and discussion, and conclusions obtained from the different sections described above.

1.3 Impact on sustainability, ethical-social and diversity

The UOC includes in all programmes the interdisciplinary competency of ethical and global commitment (CCEG), which is defined as follows: acting in an honest, ethical, sustainable, socially responsible, and respectful of human rights and diversity, both in academic and professional practice, and design solutions to improve these practices.

Thus, three main dimensions are considered: sustainability, ethical behaviour and social responsibility, and diversity and human rights. These three dimensions are aligned with the **Sustainable Development Goals** (SDG) defined by the United Nations for 2030 (Figure 1).

Figure 1. Sustainable Development Goals for 2030.

In this regard, the contributions of this project to each dimension are detailed below.

1.3.1 Sustainability

This work could contribute to scientific progress in artificial intelligence-based medical image diagnosis, including malaria and other diseases related to blood. In terms of sustainability, the project could be related to **SDG 9** (industry, innovation, and infrastructure), as it is based on a disruptive technology as the AI is, and it involves some technological advance and research and innovation to contribute to find a solution to a challenging problem, which affects in particular developing countries.

According to the WHO (1), this project could also suppose a small contribution to mitigate the effects of climate change; thus, the **SDG 13** (climate action) is involved. Temperature, rainfall, and humidity alterations could lead to changes in malaria transmission intensity, as they affect mosquito survival and parasite development within the mosquito. Having a faster diagnosis would help to improve the climate change effects related to this disease.

1.3.2 Ethical behaviour and social responsibility

Regarding the ethical behaviour and social responsibility, this work could also have a positive impact related to the **SDG 3** (Good health and well-being), as its final application is the diagnosis of malaria.

Specifically, the United Nations were committed to end the epidemics of AIDS, tuberculosis, and malaria by 2030. Nevertheless, the COVID-19 pandemic has impeded progress in SDG 3, as deaths from malaria have increased compared to pre-pandemic levels due to a decrease in vaccination and difficulty to healthcare access (6).

1.3.3 Diversity and human rights

In terms of diversity of gender, this application has not sex discrimination, as malaria affects both women and men. In contrast, it would have a positive impact in **SDG 10**, reducing inequalities based on ethnicity and income, as 94% of all malaria cases occur in sub-Saharan Africa in 2022 (1).

1.4 Objectives

1.4.1 General objectives

The two main objectives of the project are:

1. Develop a computer vision and machine learning pipeline for automatic malaria detection in peripheral RBCs.
2. Identify the RBC key image features that facilitate the analysis and interpretation of the presence of Plasmodium parasites.

1.4.2 Specific objectives

To achieve the general objectives, the following specific goals were defined:

1. Annotate RBCs manually from biomedical images of peripheral blood taken under a microscope using a specific tool.
2. Develop a pipeline for automatic segmentation of RBCs using the novel segment anything model (SAM).
3. Extract different types of image features from the segmented RBCs images.
4. Train and test a classification model for malaria detection using a classical machine learning algorithm.
5. Identify the most important features in the images that distinguish infected RBCs from uninfected ones.
6. Perform an object detection model to automatically locate RBCs from microscope images.

1.5 Approaches and Methodology

The proposed methodology in this work is based on automatic detection of RBCs infected with a parasite in blood smears based on features that differentiate infected and uninfected cells. Original images were obtained from an open-source dataset of malaria infected thin blood smears of the Hospital clinic of Barcelona (7).

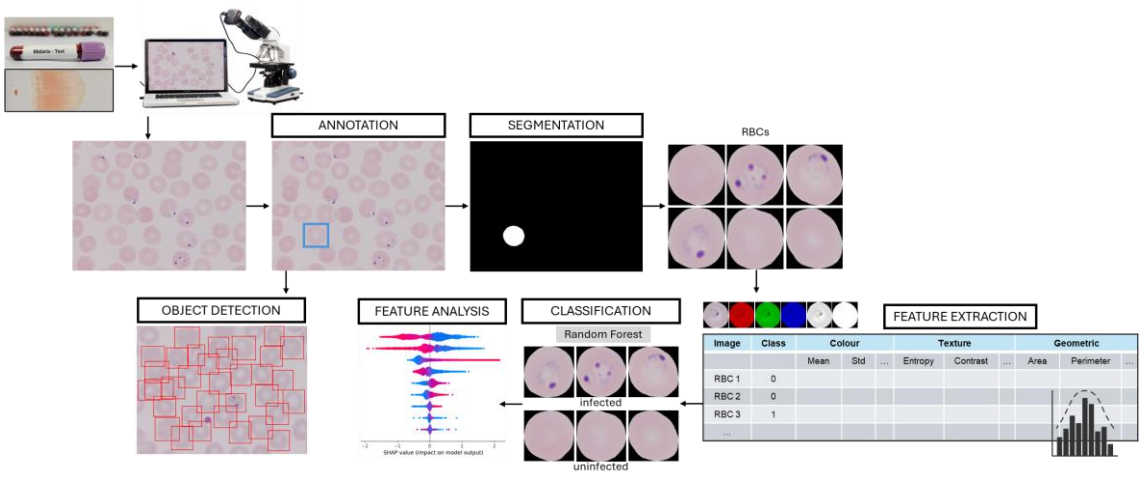


Figure 2. Main steps followed for automated malaria diagnosis.

The proposed approach includes the following steps:

1) **Annotation**: RBCs from microscopy images are manually annotated with bounding boxes. This manual annotation allows for not only continuing with the following steps to segment the cells and train the model, but also training an object detection model. Currently, there are numerous tools available for image annotation. For this work, Label Studio was used.

2) **Segmentation**: apply the SAM using bounding boxes as prompts to automatically obtain all RBCs of an image, removing artefacts and other types of blood cells. SAM model was chosen for segmentation as it is a new open-source model released in 2023 by Meta AI (5). Although its performance on natural images is impressive, here it was checked its ability to segment cells from medical optical microscopy images.

In previous works (3), RBCs segmentation methods for malaria diagnosis are mainly based on thresholding, such as Otsu algorithm, combined with morphological operations and watershed transformation. These techniques are highly used due to their simplicity, although this may not be because of their superior performance (3). Deep learning techniques have been also applied for RBCs segmentation (7). Up to our knowledge, no bibliography applying SAM model in malaria images was found.

3) **Feature extraction:** extract the most characteristic features of the previously segmented RBCs images, describing the appearance of infected and uninfected cells. Generally, three types of features are used in malaria images in the literature: colour, texture, and geometry (8,9). As parasites are stained, colour features are the most used in the bibliography to differentiate the two types of cells, as well as texture characteristics, to describe the spatial patterns of colour intensity (10). In principle, less useful in this approach, which aim is only determine the positivity or negativity of disease, morphologic features could be applied to describe the parasites found in the inside part of the cells (11). The obtention of a dataset helps for the next machine learning classification.

4) **Classification:** this step classifies RBC images into malaria infected or not. In essence, all popular classification methods of supervised machine learning have been applied to malaria diagnosis (3). Moreover, deep learning algorithms, mainly Convolution Neural Networks (CNN), are also successfully used for malaria diagnosis, although they have an impact on long computation time (12). For this project, the classical machine learning algorithm Random Forest was selected as it is among the most frequently used approaches of supervised machine learning for malaria classification with good performance results in previous studies (3). Additionally, a key advantage of Random Forest over alternative machine learning algorithms is that it is interpretable and importance measures can be used to identify relevant features of the model (13).

5) **Interpretative analysis**: determine the contribution of each feature to the model's prediction, providing human-understandable interpretation. Feature importances and SHAP values were evaluated to explain the significance of the features. This approach has been previously studied in malaria prediction using a machine learning model based on clinical data (14). However, to the best of our knowledge, no studies have applied these techniques to explain the features extracted from malaria images.

6) **Object detection:** to automatically localize RBCs within microscopy images and draw bounding boxes around them, the popular and powerful model Faster R-CNN was used. Previously, researchers have employed Faster R-CNN to localize and classify malaria-infected cells (15).

The above description is shown as a graphical abstract scheme in Figure 2.

1.6 Work Planning

The master's degree thesis project load is 15 ECTS credits, which corresponds to 375 hours of work, as each ECTS credit point can equal to 25 hours of study. The course starts on February 28th and finishes between June 25th and July 5th, depending on the date when the public defence occurs. Thus, the weekly workload is around 22 hours of work.

1.6.1 Tasks and milestones

A table with the five main tasks is shown in Table 1:

Table 1. Start and end dates of the five derivable tasks.

PEC	Start date	End date
PEC1	28/02/2024	19/03/2024
PEC2	20/03/2024	23/04/2024
PEC3	24/04/2024	28/05/2024
PEC4	29/05/2024	18/06/2024
PEC5	25/06/2024	05/07/2024

The different tasks and subtasks and the key dates at which milestones have to be hit are shown in Table 2.

Table 2. List of tasks and subtasks, their milestones, and duration in days.

Tasks	Start date	End date	Duration (days)
PEC 1. Work planning	**28/02/2024**	**19/03/2024**	**21**
Topic definition	28/02/2024	28/02/2024	1
Set objectives	28/02/2024	01/03/2024	3
Work planning definition	01/03/2024	03/03/2024	3
Literature review	01/03/2024	12/03/2024	12
Follow-up document delivery PEC1	09/03/2024	19/03/2024	11
PEC 2. Development stage 1	20/03/2024	23/04/2024	35
Data obtention and exploration	20/03/2024	20/03/2024	1
Annotation of red blood cells	21/03/2024	24/03/2024	4
Automatic segmentation (SAM)	25/03/2024	05/04/2024	12
Features extraction	06/04/2024	09/04/2024	4
Classification model (ML)	10/04/2024	20/04/2024	11
Follow-up document delivery PEC2	12/04/2024	23/04/2024	12
PEC 3. Development stage 2	24/04/2024	28/05/2024	35
Interpretative analysis	24/04/2024	14/05/2024	21
Object detection model	10/05/2024	25/05/2024	16
Follow-up document delivery PEC3	14/05/2024	28/05/2024	15
PEC 4. Report and presentation	29/05/2024	18/06/2024	21
Documentation	29/05/2024	13/06/2024	16
Presentation design	09/06/2024	18/06/2024	10
PEC 5. Public defense (1 day)	25/06/2024	05/07/2024	1

1.6.2 Calendar

The work progression during the semester is shown in a Gantt chart in Figure 3. An online tool to create the Gantt chart was used (https://www.onlinegantt.com/).

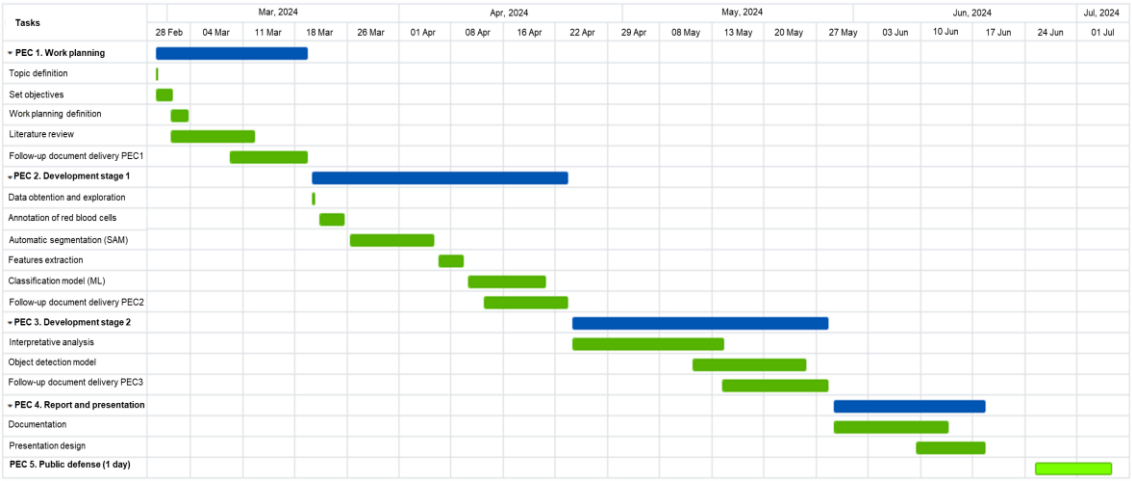


Figure 3. Gantt chart of the project planning.

1.6.3 Risk analysis

Initially, no significant risks associated with this project are foreseen. The data used for the project development was open source and already available. Additionally, the different architectures to implement were well known by the community and the advisor had previous expertise in the topic.

However, working on a laptop without good hardware could be limiting, as machine learning was involved in this project. To overcome this issue, Google Colaboratory was used to execute Python code in the cloud. Google Colab was chosen to write and execute tasks as it provides access to powerful computing resources without the need for expensive hardware. Specifically, this platform allows users to choose the best processor option based on their specific requirements. Whereas CPU is like the processor of laptops and is used for general-purpose tasks as data manipulation, GPU is a specialized processor particularly useful for simultaneous computations as those performed in machine learning, making them significantly faster than CPU.

During the time tasks were carried out, some unexpected technical problems related to the computer equipment could occur. To minimize the consequences, Google Drive was used as cloud storage for the partial deliveries and the final written report documents.

Personal problems could also become. Thus, the organization planning could be a bit flexible if this happens to finish the tasks before the delivery date. To minimize consequences and meet the deadlines, start writing from the very beginning was also planned.

1.7 Brief summary of the obtained products

Many different products will be obtained at the end of the project:

- Different partial deliveries (PEC), including a working plan.
- Final written report, which includes an introduction, objectives, results and discussion, and conclusions.
- A video that shows a concise oral presentation with a graphical support explaining the most notable parts of the project.
- GitHub repository with the code: https://github.com/mursune/TFM_malaria.

1.8 Brief description of the other chapters of the report

- Chapter 2 includes the **state of the art** of malaria diagnosis with traditional methods and AI.
- Chapter 3 describes the **methodology** used.
- Chapter 4 shows the **results and discussion**.
- Chapter 5 explains the final **conclusions** of the project.
- Chapter 6 lists some **relevant words** or concepts.
- Chapter 7 includes the **bibliography**.
- Chapter 8 contains the **appendix.**

2 State of the art: malaria diagnosis

2.1 Traditional methods

Several methods of malaria diagnosis exist worldwide. However, diagnosis with **optical microscopy** is the gold standard and it is chosen in most cases.

Two types of blood smears can be distinguished in malaria diagnosis depending on the thickness: thick or thin (Figure 4). A **thick blood smear** is a drop of lysed blood on a glass slide, and it allows to examine a larger sample of blood. Thus, thick smears are better to use as a first test for a positive/negative diagnosis when often a few parasites are present in the blood, with more sensitive by 30 times. In contrast, a **thin blood smear** consists of a drop of blood spread across a glass slide, allowing the identification of the parasite species and development stage as well as obtaining a quantification of malaria parasitaemia (16). The microscopic slides are examined with a 100x oil immersion objective. For a negative diagnosis, it should be inspected a minimum of 100 Field of View (FOV) for thick films or 800 FOV for thin films (4).

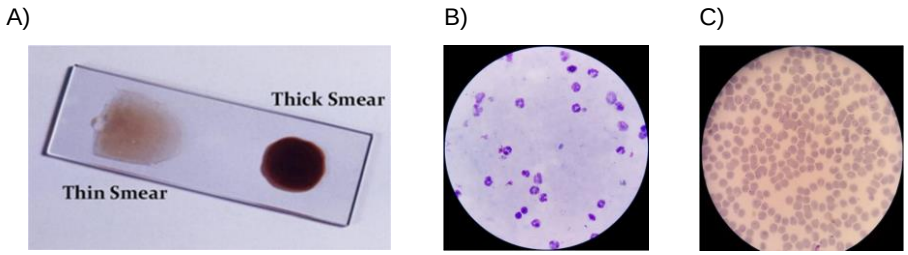


Figure 4. A) Preparations of thick and thin blood smears (17). (B) Microscopic images of thick and (C) thin smears. (18)

The diagnosis of malaria using optical microscopy is intricately linked to understanding the life cycle of the malaria parasite (Figure 5), particularly within the human host. In humans, after a first phase where parasites infect liver cells and replicate (exo-erythrocytic cycle), they invade the bloodstream and multiply in the erythrocytes (erythrocytic cycle). Diagnosis occurs during this second phase, where four different stages of the parasite can be found (rings, trophozoites, schizonts and gametocytes). These morphologies of infected RBCs and parasites are used to diagnose malaria. Additionally, the specific appearance of the parasite and cells is used for species identification. Five species of Plasmodium can infect humans: *P. falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi.* The most common species are *P. falciparum* and *P. vivax,* whereas *P. falciparum* is responsible for most malaria-related deaths.

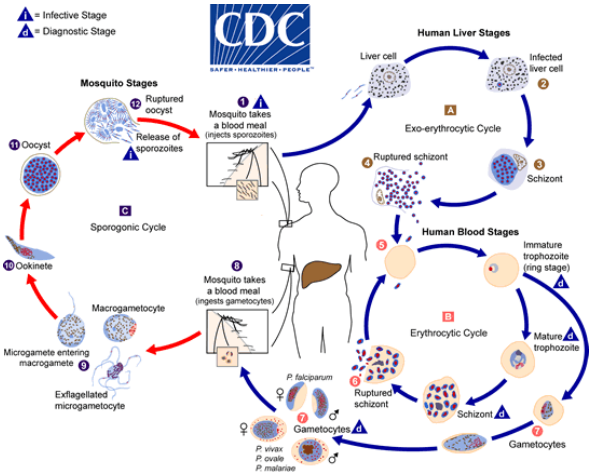


Figure 5. Life cycle of malaria (19).

Fluorescence microscopy has the potential to improve malaria diagnosis. Acridine orange is the most common dye employed, which is excited at 490nm and displays green fluorescence (20). The limitation of fluorescence microscopy is that it requires expensive equipment, and it cannot easily differentiate the diverse species of parasites. Moreover, the dye is nonspecific, as it stains nucleic acids from all cell types in the sample.

As an alternative technique to microscopy, **rapid diagnostic tests (RDTs)** are used. These easy handling and fast tests can detect antigens of the parasite and provide results in 2-15 minutes. However, a positive diagnosis with RDT should be followed by microscope observation to quantify and confirm the species of parasite.

Molecular diagnosis using polymerase chain reaction (PCR) is also available to detect DNA of the parasite. It allows higher sensitivity and specificity than conventional microscopy or RDTs and can identify the species of parasites (21). However, it requires good laboratory infrastructure, expensive equipment and is not usually adequately implemented in developing countries (22).

Serology techniques to detect antibodies against malaria parasites also exists. However, indirect immunofluorescence (IFA) or enzyme-linked immunosorbent assay (ELISA) can only detect the exposure to the parasite, but not the current infection.

2.2 Artificial intelligence methods

Novel diagnostic strategies to fight against infectious diseases have been developed in recent years, including techniques of automatic image analysis based on **artificial intelligence (AI).** Therefore, AI has a high impact on health, as it helps healthcare professionals make decisions related to clinical diagnosis and treatment.

In the case of malaria diagnosis, these alternative tools emulate the microscope visualization by experts of blood smear samples and automate the procedure, resulting in faster and low-cost diagnostics. Likewise, the diagnosis of malaria would require less supervision and could be more reliable, as human error in traditional microscopy is identified to be a major cause of inaccurate diagnostics (4).

Both classical machine learning and deep learning algorithms have been tested for malaria diagnosis in thick and thin blood smear digital images, resulting in notable improvements compared to traditional methods due to their faster procedure.

2.2.1 Machine learning

Machine learning can provide computer with the ability to learn the relationship between the input (training data) and the output using several algorithms, and it can then predict the output with new data. The most used technique of machine learning is **supervised machine learning**, which is characterized because it uses labelled datasets to train algorithms to predict outputs.

Supervised learning is divided into two main categories depending on the type of output: regression, if the outputs are continuous values, or classification, if the predictions are discrete. Classification can be binary, if there are only two possible values (such as yes or no, infected or uninfected, true or false, etc.), or multiple classifications, if the number of outputs is more than two. Some common techniques used in supervised machine learning to solve classification problems are logistic regression, decision tree, random forest, support vector machines (SVM), k-nearest neighbours (KNN) and naive Bayes.

Specifically for the diagnosis of malaria, the most used algorithm for image classification in the literature is SVM (3). Although it generally outperforms all other classifiers with higher classification accuracy (23), its main drawback is that it is less interpretable compared to decision trees. An exception is shown in (24), where Random Forest system achieves better accuracy than SVM.

2.2.2 Deep learning

Deep learning employs artificial neural networks with multiple layers to learn, extracting automatically relevant features from the data. The most widely used method is the **Convolutional Neural Network (CNN)**, a subset of deep learning that is designed to learn large volumes of images for classification. Highly effective for larger datasets with complicated patterns, in general, it is often preferable to traditional machine learning, as it needs less human intervention. A simple scheme to illustrate the position of CNNs within the field of AI is shown in Figure 6.

For malaria diagnosis, various common architectures of deep learning have been tested with great performance to classify images into parasitized or non-parasitized, including custom-built CNN, VGG, MobileNetV2, ResNet, AlexNet, EfficientNet and DenseNet (7,25–27).

Although deep learning has good results in detecting malaria, a negative point is that large training sets are typically needed, but annotated training images in medical applications are not easy to obtain due to the requirement of expert knowledge and privacy concerns (3). Moreover, these models are complex and can be difficult to interpret. Likewise, training deep learning models is computationally intensive and requires powerful hardware (GPUs). This need for more sophisticated hardware components makes it less affordable, which is critical since most malaria-endemic regions are resource-limited areas. Novel approaches based on deep learning to achieve low-cost diagnosis using smartphones are being developed (28,29).

Deep learning-based models have also been successfully applied for **object detection**. One of the top-performing object detection models in recent years is Faster Region-based Convolutional Neural Network (**Faster R-CNN**), which has demonstrated good performance not only in natural images, but also in detecting erythrocytes and leukocytes from microscopy blood images (30).

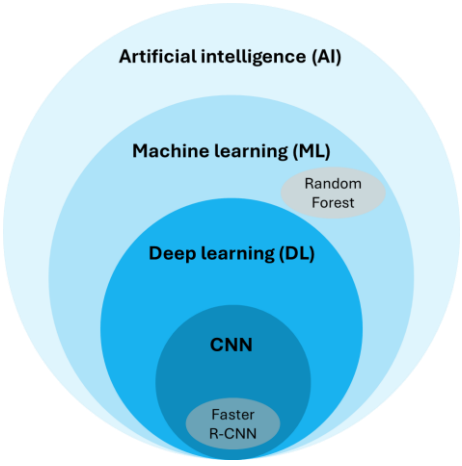


Figure 6. Scheme of artificial intelligence organization.

This work aims to perform an analysis of optical images from thin blood smears infected by malaria parasites. The method is based on feature extraction and analysis from previously segmented RBC images with the novel SAM model. The Random Forest algorithm is chosen as the algorithm as it allows to investigate the features importance. The process could be automatized by detecting RBCs with Faster R-CNN. The study would provide the scientific community other options for malaria diagnosis and more understanding about important features in the classification of infected and uninfected images, thus, the motivation of this work.

3 Materials and methods

3.1 Tools

Google Colaboratory was used to carry out this work. It is a web platform developed by Google that allows people to edit and run Jupyter Notebooks directly from the web, only with a Google account.

The main **advantages** of Google Colab are:

- It is not necessary to have anything installed, since all code runs on the servers of Google.
- It allows you to connect to Google servers that have GPUs.
- Simple connection to Google Drive, where data can be saved.
- The project can be shared and edited with other Google accounts.

Google Colab has also some **disadvantages**:

- The duration of sessions is limited. After a certain period of inactivity, the session may disconnect, which is inconvenient for long-running tasks.
- Storage space for notebooks and data is limited. This is an obstacle when working with big image datasets, as it causes high data usage.
- Although there is free GPU access, its access depends on demand and it can be limited.
- It requires a stable Internet connection, as it operates in the cloud.

To solve the issues of limited session duration and GPUs access, Google Colab Pro was employed.

3.2 The dataset

A total of 331 images of thin blood smears from five malaria patients obtained in the Hospital clinic of Barcelona were used in this work. They are RGB images with a size of 2400x1800 pixels in JPG format, acquired using an Olympus microscope BX43 with 1000x magnification and an Olympus camera DP73. The dataset is available in (7).

To obtain the images, peripheral blood samples from patients were collected in EDTA tubes as anticoagulant. Then, thin peripheral blood smears stained with May Grünwald-Giemsa were obtained using the Sysmex sp-1000i automated slide preparer stainer.

3.3 Annotation

Images from thin blood smears contain three different types of cells: platelets, white blood cells, and RBCs. Only the last ones are of interest for the diagnosis of malaria as they are the ones infected by parasites.

Erythrocytes from the images were labelled using the software application **labelstud.io**. (https://labelstud.io/). This platform allows to select every erythrocyte from the images and manually label them as uninfected RBCs (in blue) or malaria-infected RBCs (in red). An example is shown in Figure 7.

Object detection with bounding boxes as labeling setup was used. Bounding boxes are rectangular frames drawn around the cells that can be defined by four coordinates, which are the x and y of the top-left corner and the bottom-right corner. These bounding boxes allow to identify and localize each cell in the image.

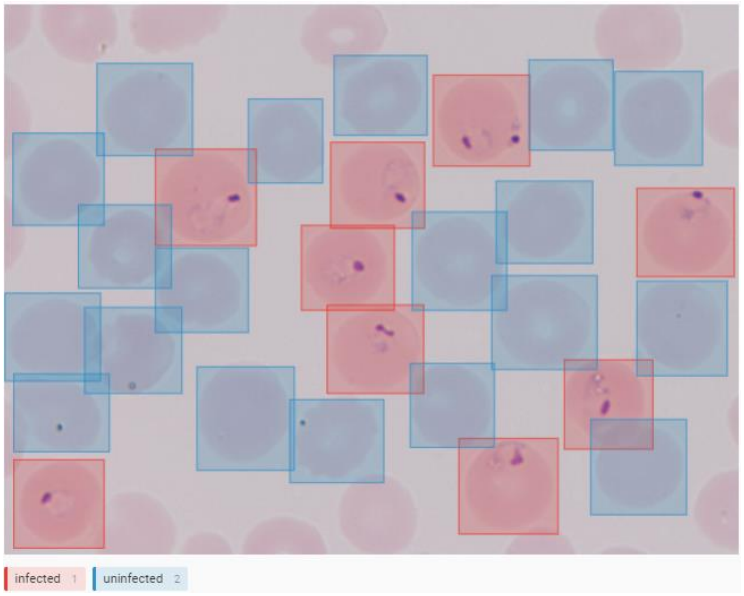


Figure 7. Image annotation in Label Studio

The bounding box annotations were exported as COCO format, with a JSON file and a folder with the original images. The JSON file allows an efficient storage of bounding box annotations, and it includes information about:

- Images: contains a list of images and information about each image, such as id, width, height, and file name.
- Annotation: contains the coordinates (in pixels) of the bounding box and the total area of this rectangle around the object. It also includes the image id and the category label.
- Categories: contains the id of each category (0 or 1) and the name of the category (uninfected or infected).

3.4 Segmentation

The bounding boxes generated by Label Studio provide a general indication of the location of each RBC within the images. However, these rectangles often include background regions, portions of other objects, or other types of cells that are not relevant to the analysis. To separate the object of interest from other elements, drawing its exact shape, a segmentation step is needed.

Segment anything model (SAM) was applied for segmentation in this work. It is a new open-source tool for automatic image segmentation that was released in April 2023 by Meta AI. It is a powerful zero-shot foundational model. Specifically, it is based on Transformer vision models, an adaptation of the transformer architecture for computer vision tasks, a type of deep learning architecture which originally were design for natural language processing. SAM was trained on a diverse dataset (called SA-1B) of over 1.000 million masks for image segmentation that can generate masks for images and objects that it has not seen during training.

The model results in some advantages in image segmentation, as it is:

- Promptable: can find a mask given different types of prompts such as points, bounding boxes, text, or a combination.
- Ambiguity-aware: the model returns not only one mask, but the most likely options. This characteristic allows to face ambiguity about the object to be segmented, solving a problem for image segmentation in the real world.

In this study, segmentation masks are created with SAM using bounding boxes as prompts. A comparison between obtaining one or three masks as output was carried out.

The **model architecture** of SAM, which is represented in Figure 8, has three fundamental components:

- **Image encoder:** Mask Auto-Encoder pre-trained Vision Transformer (ViT) is used. It is an adaptation of the ViT capable of processing high resolution images. The model can work with three different encoders: ViT-B, ViT-H and ViT-L. The encoder ViT-H was chosen for this work, as it is substantially better than ViT-B, but does not have great improvements compared to ViT-L, although the latter has a larger size. The output of the encoder is an embedding of the input image reduced 16 times, typically from 1024×1024 of the input image to 64×64. This downsizing process is crucial for efficient processing while retaining essential image features. It is carried out only once per image, so, it does not depend on the prompt.
- **Prompt encoder:** There are two types of prompts: sparse (points in the image, bounding boxes or text) or dense (segmentation masks).
- **Mask decoder:** A modification of the transformer decoder block is used to predict masks. The decoder considers the embeddings of the prompt and the image. The decoder outputs an embedding, which is then mapped to a linear classifier. This classifier predicts the probability that a given point belongs to the mask. To solve ambiguities, a maximum of three possible masks with its own confidence score can be obtained.

A Python environment in Google Colab with access to a GPU was used for faster processing. After the segmentation step, the masks of all previously annotated cells with bounding boxes were obtained.

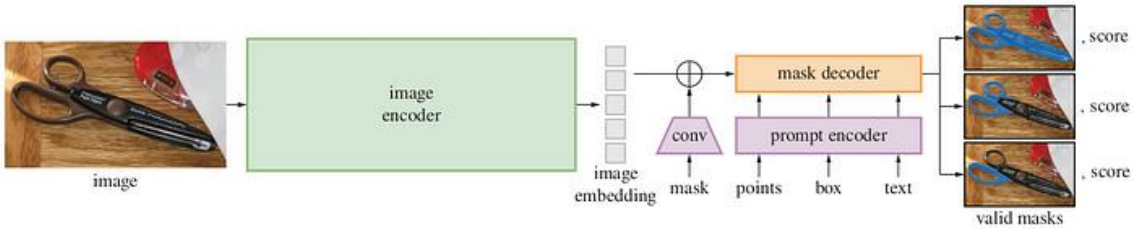


Figure 8. Segment Anything Model (SAM) overview (5).

3.5 Feature extraction

Feature extraction is the process of obtaining relevant and discriminative properties from images to transform the higher dimensional data to fewer dimensions, and this, reducing complexity of analysis. It constructs the combination of variables in such a way that it reduces the resource usage and describes the data with sufficient accuracy.

The selected features to extract from RBC images in this work are shown in Table 3.

Table 3. Summary of the colour, texture, and geometric features extracted

Type	Channels	Features
Colour	RGB	Mean, Standard deviation, Kurtosis, Skewness, Histogram
Texture	Grayscale	Mean, Standard deviation, Kurtosis, Skewness
	Grayscale and RGB	Entropy, Contrast, Correlation, Energy, Dissimilarity, Homogeneity
Geometric	Binary	Area, Perimeter, Circularity, Eccentricity

3.5.1 Colour features

Five different features related to colour were extracted from each cell image of infected and uninfected classes. The colour features were computed separately for each colour channel (red, green and blue in RGB colour space) to obtain more information. The chosen colour features were the following:

- **Mean colour:** measures the average colour of the image for each colour channel.
- **Standard deviation colour**: represents the variation of colours in the image. It provides information about the diversity of colours present in the image.
- **Kurtosis**: measures the "tailedness" of the intensity distribution of each channel, indicating whether the distribution is peaked (positive value) or flat (negative value) compared to a normal distribution.
- **Skewness:** measures the asymmetry of the intensity distribution of each channel around its mean. A positive value indicates a long tail to the right of the mean, while a negative value indicates a long tail to the left of the mean. Both kurtosis and skewness provide information about the shape of the intensity distribution of the colour channels in an image.
- **Colour histogram**: represents the frequency of occurrence of different colour values across each colour channel.

The library `NumPy` was used to obtain mean, standard deviation, and colour histogram. Kurtosis and skewness were calculated with `spicy.stats`, which contains many statistical functions.

3.5.2 Texture features

The evaluated texture features of the images were:

- For grayscale images, the mean, the standard deviation, the kurtosis and the skewness were calculated as for RGB images.
- **Entropy:** can be used to evaluate the texture and complexity of an image. Higher entropy indicates more complexity and randomness (high detailed or noisy image), while lower entropy indicates less complexity (uniform or smooth images).
- Features derived from the **Gray Level Co-occurrence Matrix (GLCM)**, which considers the spatial relationship of pixels. These features are:

- **Contrast:** measures the intensity contrast between a pixel and its neighbours. High contrast values indicate intensity differences between neighbouring pixels, implying more texture. Contrast emphasizes larger differences.
- **Correlation**: measures how correlated a pixel is with its neighbour.
- **Energy**: measures texture uniformity. High energy values correspond to homogeneous textures, with little variation in intensity values.
- **Dissimilarity**: measures the difference between pairs of pixels. It is more influenced by small and medium differences than by very large differences as contrast.
- **Homogeneity:** higher homogeneity values indicate a more uniform texture.

The library `scikit-image` was used to obtain entropy values with `shannon_entropy`() function and features derived from the GLCM with `greycomatrix()` and `greycoprops()`.

3.5.3 Geometric features

The geometric features calculated were:

- **Area:** is the number of pixels contained within the cell, indicating its size.
- **Perimeter:** measures the boundary length of the cell.
- **Circularity:** measures how close the shape of the cell is to a perfect circle. A perfect circle has a circularity value of 1 and lower values indicate more elongated or irregular shapes. It is calculated by the following formula:

$$\text{circularity} = \frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$$

- **Eccentricity:** is a dimensionless parameter that ranges from 0 (a perfect circle) to 1. It is a measure of how much an ellipse deviates from being circular.

Values of area, perimeter, and eccentricity were calculated from a binary image with the function `regionprops()` from the library `scikit-image`. Circularity was obtained from area and perimeter by applying the formula.

After the feature extraction step, a dataset with the described features above was obtained. Moreover, for the following machine-learning training purposes, each group of features was labelled according to the previous annotation. Not only the corresponding labels but the paths of the RBC images were added to check the misclassified images in the model. The structure of the dataset is shown in Figure 9. This quantitative information provided by feature extraction is useful for subsequent learning and classification.

Image	Class	Colour features			…	Texture features			…	Geometric features			…	Image path
		Mean	Std	Kurtosis		Entropy	Contrast	Correlation		Area	Perimeter	Circularity		
RBC 1	0													
RBC 2	0													
RBC 3	1													
RBC 4	0													
RBC 5	1													
…														

Figure 9. Structure of dataset, which includes features, class and image path for each image.

3.6 Machine learning classification

3.6.1 Random forest algorithm

Once the features were extracted from images, they were used to train a machine learning model to perform image classification.

The algorithm Random Forest can be used not only for classification but also for regression problems. The model is called ‘random forest’ as it consists of a multitude of decision trees during training, and it is built taking a random sample of the data. It is based on the concept of **ensemble learning**, which is a process of combining multiple classifiers to solve a complex problem.

A representation of a Random Forest model is illustrated in Figure 10. Each tree is a simple model that splits the data into a random subset of features. It creates a tree-like

structure where each node in the tree represents a feature from the input space, each branch a decision and each leaf at the end of a branch the corresponding output value. The output for classification tasks is the class of most of the trees, whereas in regression tasks, the prediction is the average of the outputs from all the individual trees. The aim of this work is to classify images in two classes: infected and uninfected.

The training algorithm for random forests applies the general technique of **bagging** or bootstrap aggregating, as it selects a random sample with replacement of the training set and fits trees to these samples. This helps in reducing variance and avoiding overfitting.

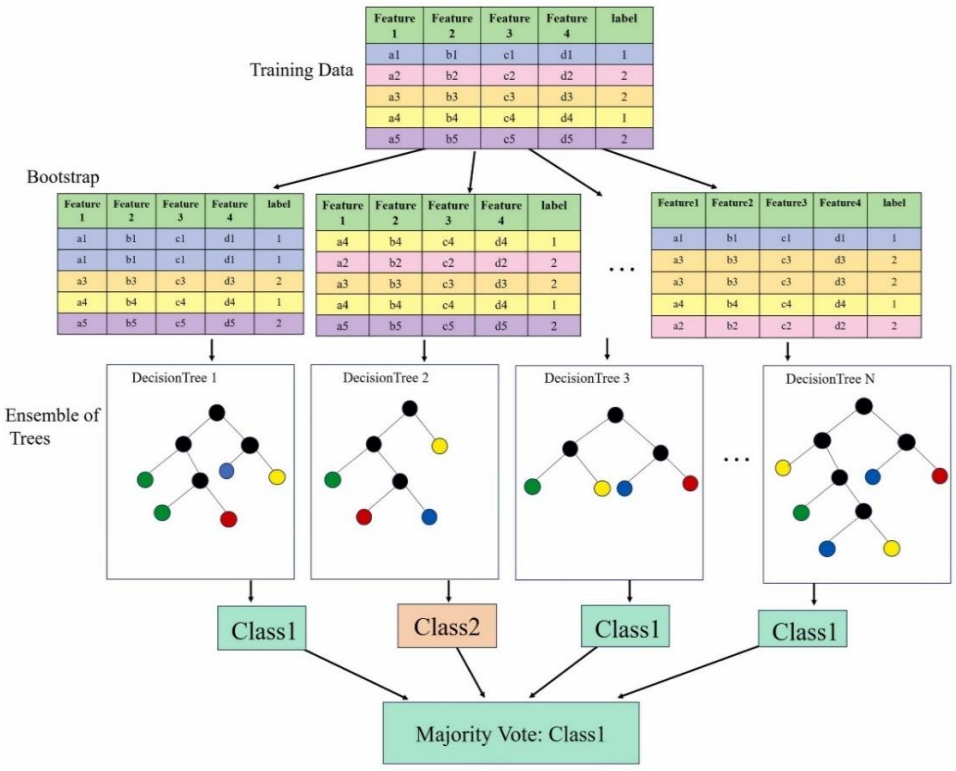


Figure 10. Random forest scheme (31).

Then, a Random Forest classifier was created using the `BalancedRandomForestClassifier` class and 10-fold cross-validation. The object `StratifiedKFold` from the library `sklearn` provides train/test indices to split data in train/test sets. It is a variation of `KFold` that ensures each fold preserves the percentage of samples for each class as in the original dataset.

Finally, the model was evaluated obtaining the **confusion matrix** (Figure 11), which includes:

- **True Positive (TP):** number of instances that are correctly predicted as positive.
- **True Negative (TN):** number of instances that are correctly predicted as negative.
- **False Positive (FP):** number of instances that are incorrectly predicted as positive. Also known as Type I error.
- **False Negative (FN):** number of instances that are incorrectly predicted as negative. Also known as Type II error.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 11. Confusion matrix.

From the confusion matrix, the following metrics are calculated:

- **Accuracy:** the proportion of correctly classified instances (both positive and negative) out of the total instances. It is calculated as:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precision:** the proportion of correctly predicted positive instances out of all instances predicted as positive. It is calculated as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + +\text{FP}}$$

- **Recall:** the proportion of correctly predicted positive instances out of all actual positive instances. Also known as Sensitivity or True Positive Rate (TPR). It is calculated as:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + +\text{FN}}$$

- **F1-score:** The harmonic mean of precision and recall, providing a balance between the two metrics. It is a useful metric for binary classification tasks, especially when the classes are imbalanced. It is calculated as:

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

3.7 Features evaluation

For features evaluation, a Random Forest model with undersampling and balancing was created. In this case, the dataset was split into training and testing sets using the `train_test_split` function, with the 80% of data for training and 20% of data for testing, a typical partition of datasets in bibliography. A `random_state=42` was indicated to ensure the same split each time.

3.7.1 Feature importance: Gini

Feature importance in a random forest model consists of assigning a score to input features based on their significance in predicting a class. It is useful to identify which features contribute the most to the model and can be important for feature selection and model interpretability.

The feature importance can be determined by the method of **Mean Decrease Impurity (MDI)**, also known as **Gini importance**. This method is based on the total decrease in node impurity brought by a feature across all the trees in the forest.

When a feature is used to split a node in the decision tree, the impurity is reduced. The reduction in impurity is a measure of how well that feature splits the data. The total importance for a feature is calculated as the sum of the impurity reductions it contributes

across all the nodes where it is used to split the data, averaged over all trees in the forest. Thus, a feature with high Gini importance has a substantial impact on reducing impurity, indicating it plays a crucial role in the model's predictions.

Features importance of the model were calculated with the attribute `model.feature_importances_` from the `scikit-learn` library.

3.7.2 SHAP values

For a more comprehensive understanding of feature importance, SHAP values were calculated, as well.

SHAP values approach, from SHapley Additive exPlanations, is based on **cooperative game theory**, specifically on the concept of Shapley values. The Shapley value, introduced by Lloyd Shapley in 1953, is a method to fairly distribute the total gains (or costs) among the players based on their individual contributions. SHAP values apply this concept to interpret the contributions of features in a machine learning model's prediction, considering the model prediction as a cooperative game where features are the players.

SHAP values can be applied to any machine learning model, unlike feature importance. They measure the impact of each feature on model predictions by evaluating the change in the predicted outcome when including or excluding that feature. Positive SHAP values indicate features that contribute to increasing the prediction, while negative values indicate features that decrease the prediction. In addition, SHAP can help identify not only which features are important but also how they influence specific predictions.

SHAP values were calculated with `shap.TreeExplainer` provided by the `shap` library, which is optimized for computing SHAP values for tree-based models. First, the TreeExplainer was initialized with the Random Forest model. Then, the TreeExplainer object was used to compute Global SHAP values for the test data. Global SHAP values for the entire dataset and per class were visualized using summary plots.

Local SHAP values were also calculated with `shap.force_plot` in 10 specific instances (5 for uninfected samples and 5 for infected samples). This is particularly useful for understanding a particular prediction of the model for a specific instance.

3.8 Object detection model

An object detection model was created to automatically detect RBCs in field-of-view images and generate potential bounding boxes. Thus, this model enables the automation of the analysis of blood images obtained from microscope. After identifying the bounding boxes using the object detection model, segmentation with SAM and classification with a Random Forest machine learning model complete the automation of sample analysis for malaria diagnosis.

A Faster R-CNN pretrained object detection model from `torchvision` was selected due to its widespread use in the scientific community for object detection tasks. Its architecture is based on a standard CNN, to extract deep feature maps from the input image, and a Region Proposal Network (RPN), a small neural network that generates a set of bounding boxes candidates and assigns scores to these proposals based on their likelihood of containing an object.

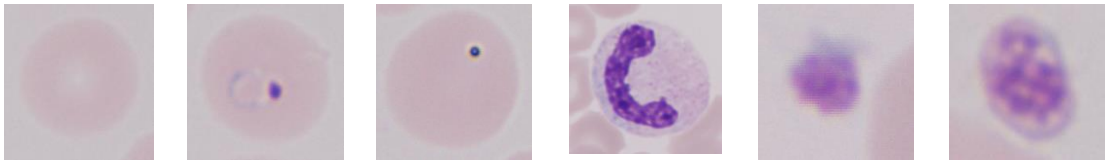
Annotations obtained from LabelStudio, along with images, were used to train the object detection model and identify bounding boxes around cells. Then, the model was fine-tuned for our number of classes, that in our case is 2: one class representing RBCs in general (undifferentiated by subclass) and another class representing the background.

4 Results and discussion

4.1 Annotation

All RBCs from images were annotated with Label Studio using bounding boxes, labelling them as infected or uninfected.

Different types of cells were found in the original images apart from RBCs, as white blood cells, platelets and gametocytes (Figure 12). These objects were not labelled for this study. In addition, it is worth highlighting that some RBCs with stained cytoplasmatic inclusions, called Howell-Jolly bodies, can be easily confused with infected RBCs, although they must be labelled as uninfected.



Uninfected RBC | Infected RBC | RBC with Howell–Jolly bodies | WBC | platelet | Gametocyte

Figure 12. Example of the different type of cells found in the samples.

The results of annotation were visualized in Google Colab to check the upload was properly done. An example of annotated image is shown in Figure 13.

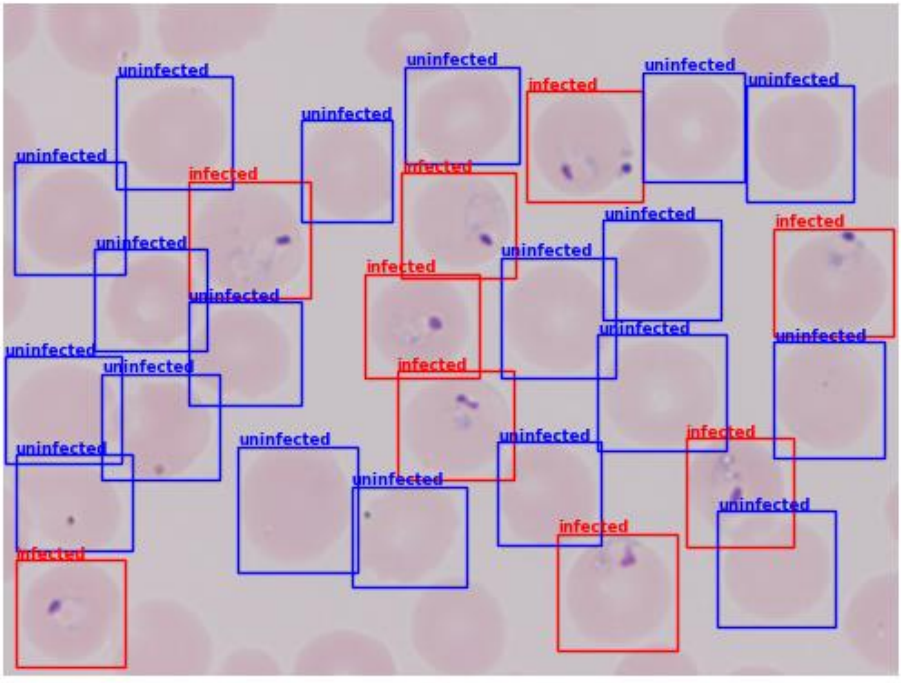


Figure 13. Annotated image with LabelStudio visualized in Google Colab.

4.2 Segmentation

After annotation of all original images with Label Studio, individual RBCs were needed for further steps. SAM was tested for segmentation.

Initially, the multimask_output = False option was tried, obtaining only one mask option. However, some of the images were not properly segmented, especially those from patient 5, in which the centre of the cells had a lighter colour (Figure 14.A).

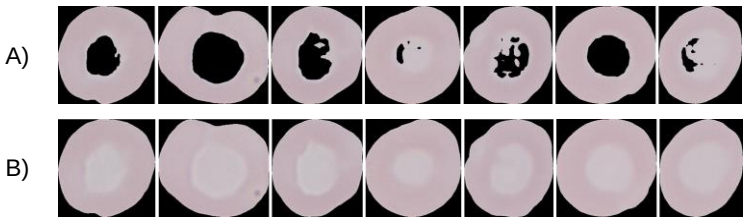


Figure 14. Comparison of segmented images from patient 5 with SAM (A) without multimask option or (B) with multimask option.

Hence, the multimask_output =True option was utilized to generate three masks as outputs with a score value of the quality of these masks. Then, the best single mask was chosen considering the one with the highest score, observing an improvement in the segmented images with SAM (Figure 14.B). An example illustrating the overlap of the three obtained masks with the original images, along with their respective score value is in Figure 15. In that case, mask 3 was the definitive due to its highest score.

Figure 15. Image overlayed with the three masks obtained with the multimask option of SAM.

Therefore, to visually verify segmentation, an overlapping of the source images and the definitive masks was done. An example is depicted in Figure 16.

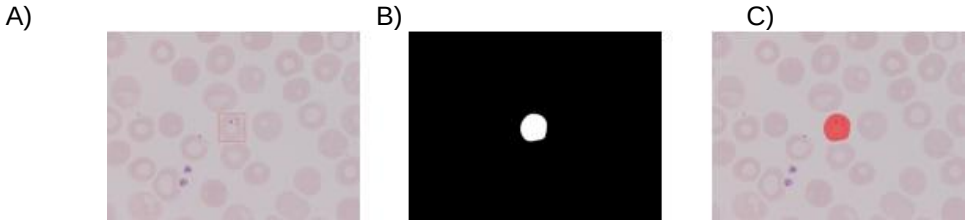


Figure 16. Visually verification of segmentation. A) source image with bounding boxes, B) definitive mask (the one with a highest score) and C) image overlapping.

After checking the obtained masks, the goal was to get each individual cell. This was accomplished by overlaying the mask and the source image using `cv2.bitwise_and` (Figure 17.A). Subsequently, the masked image was cropped obtaining the coordinates of the minimum enclosing rectangle with `cv2.boundingRect` (Figure 17.B).

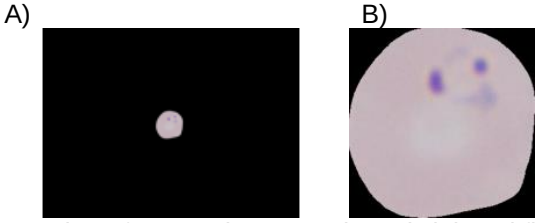


Figure 17. Overlap of source image and mask (A) and final image (B).

The final cropped images were checked manually. Overall, satisfactory segmentation performance was achieved. Some examples of infected and uninfected segmented RBCs are imaged in Figure 18.

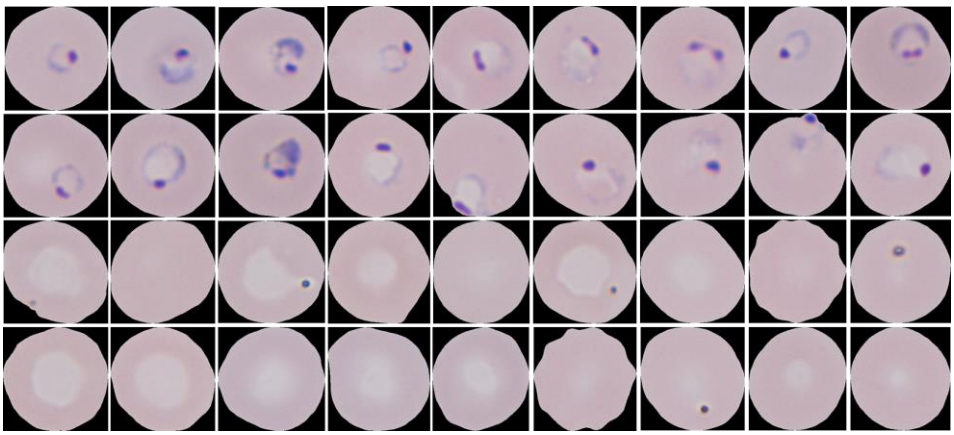


Figure 18. Example of segmented RBCs.

Although not discarded, it is interesting to consider that some uninfected RBCs appear incomplete, as shown in Figure 19. This could be explained by the overlapping of cells in high-density areas, where the underlying cell was cut, obtaining a RBC formed by only the visible part of the cell in the image. This occurrence is more likely with uninfected cells due to the imbalance in the dataset, where uninfected cells are more prevalent than infected ones.

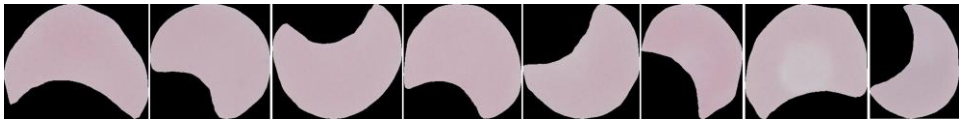


Figure 19. Example of not entire uninfected RBCs.

After all, the efficiency of SAM model to perform a segmentation of RBC images from thin blood samples was 98.3%, as 149 images out of a total of 8.750 were discarded. Some examples of discarded images are shown in Figure S1 and S2 from the appendix.

After the segmentation step, the images were organized in two separate folders based on class for further steps, having a total of 901 infected images and 7.700 uninfected images.

4.3 Features extraction

Colour, texture and geometric features were extracted. For this purpose, red, green and blue components from the original RGB images, as well as grayscale and binary images were obtained. An example of each type of image from an infected and uninfected RBC is illustrated in Figure 20.

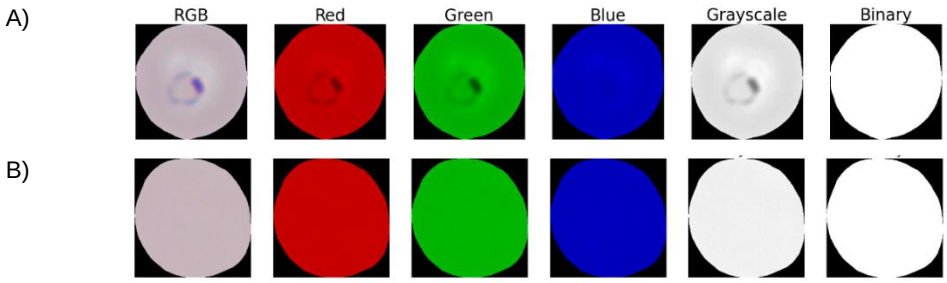


Figure 20. Example of A) infected and B) uninfected RBC images in RGB, red, green and blue components from RGB image, grayscale image and binary image.

An optimization of the threshold value to obtain binary images was performed and it is included in Figure S3 from the appendix.

Among other types of features, the histograms from red, green and blue component images are included. Histograms count, for each colour channel, the number of pixels that fall into a specified number of intensity bins. A comparison of histograms between an infected RBC and an uninfected RBC is shown in Figure 21.

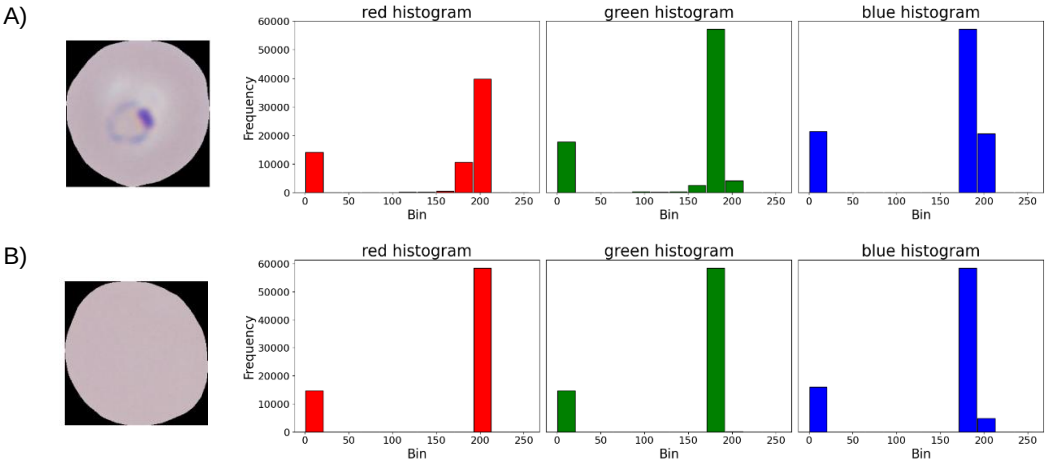


Figure 21. Example of A) infected and B) uninfected RBC image in RGB and their corresponding histograms in red, green and blue components.

The obtained dataset comprised 8.601 rows representing the total number of images and 80 columns containing features values. From the total samples, 7.700 were uninfected while 901 were infected, which correspond to 89.5% and 10.5%, respectively.

Therefore, the obtained dataset had a significantly higher number of observations for the uninfected class, while the number of rows for the infected class had a notably lower count, resulting in **imbalanced data**. This class imbalance can affect the subsequent prediction with machine learning algorithms.

4.4 Machine learning classification

The problem of having an imbalanced dataset is that the model may tend to predict the majority class with higher probability. To improve the handling of the imbalanced dataset and achieve more reliable predictions for both classes, three strategies were implemented:

- **Undersampling:** is a technique for imbalanced datasets to reduce the number of observations in the majority class, which in this case was the uninfected class.
- **Use of the `BalancedRandomForestClassifier:`** this classifier adjusts the class weights during training and helps mitigate the impact of class imbalance.
- **Evaluation metrics other than accuracy**: evaluation metrics such as precision, recall, and F1-score were considered to assess model performance.

As mentioned, a random undersampling approach was applied to remove uninfected data, with a fixed seed to ensure reproducibility of the experiments. It was decided to perform subsampling so that the infected class had approximately 30% of the number of samples of the uninfected class (before undersampling, the infected observations corresponded to 10.5%). The total number of data points and percentages before and after undersampling are shown in Table 4. The dataset after undersampling was used to train and test the model.

Table 4. Comparison of data before and after undersampling.

	Uninfected total	Uninfected %	Infected total	Infected %	Uninfected to remove	Total samples
Before undersampling	7700	89.5	901	10.5	5598	8601
After undersampling	2102	70.0	901	30.0	-	3003

After undersampling, the Random Forest model was trained with 10-fold cross-validation using 100 trees. In k-fold cross-validation, the dataset is divided into k smaller sets (in this case, k=10). For each fold, the model is trained using k-1 of the folds as training data. The resulting model is then validated against the remaining part of the data, which is used as a test set to calculate the performance metrics.

The metrics of the model performance are shown in Table 5.

Table 5. Model performance metrics

TP	FP	TN	FN	accuracy	precision	recall	F1
896	10	2092	5	0.995	0.989	0.994	0.992

An accuracy of 99.5% was achieved, with ten false positives and five false negative. In imbalanced datasets, accuracy can be misleading because the model might predict the majority class most of the time and still achieve high accuracy. To take this into account, other metrics of the model were calculated. Precision measures how many of the predicted positives are true positives, while recall (sensitivity) measures how well the model identifies all actual positives. The F1-score combines these to focus on the performance of the model concerning the minority class. In this case, F1-score was 99.2%. It should be considered that these metrics are overestimated, as they do not include mis-segmentation error.

The misclassified images (false positives and false negatives) are shown in Figure S4 of the appendix. Some of the labels of these missclassified images are not entirely clear, as the possible parasites are not properly visible, likely due to poor focus of the microscope during sample measurement.

Thus, a model with high values of accuracy, precision, recall, and F1 score was achieved, indicating that it correctly classifies almost all instances with low number of errors. Such good results might be explained by the simplicity of the problem, with images where the patterns between infected and uninfected RBCs are clear and distinct, and the classes are easily separable. While the model achieves remarkable scores, it would be essential to ensure that it generalizes well to new images from different patients.

In this work, both false positives and false negatives were low. However, their implications for malaria diagnosis differ significantly. A false positive can cause unnecessary stress to the patient and their family, as it implies that a patient is diagnosed with the disease, but they do not have it. In contrast, a false negative can delay treatment and lead to more severe health problems, as it means that patients are not diagnosed with malaria when they actually have it. F1 is a particularly useful metric when the minority class is critical, as in this case.

4.5 Features evaluation

4.5.1 Features importance

To visualize feature importance, a bar plot is shown in Figure 22, where the x-axis lists the features and the y-axis represents the feature importance scores.

Gini importance is a metric used to measure the relative importance of each feature in predicting the class. It indicates the reduction in the impurity of each feature during the training process. Higher importance values indicate that these features have a larger impact on the model's ability to make predictions. Moreover, Gini importance is useful for feature selection, as features with lower importance values might be considered less relevant and could be removed from the model without significant impact, selecting features with higher importance values.

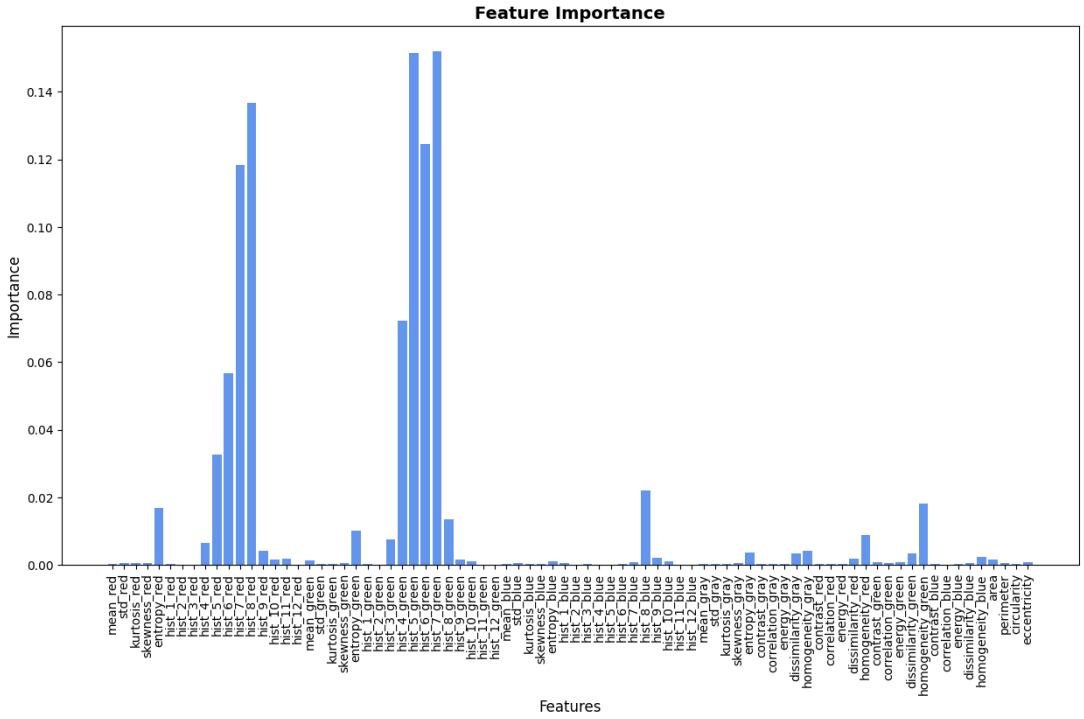


Figure 22. Features importance bar plot.

In Table 6, the 20 most important features with their importance values are shown.

Table 6. Features importance values of the 20 most important features.

position	Feature	Importance	position	Feature	Importance
1	hist_7_green	0.151905	11	entropy_red	0.016890
2	hist_5_green	0.151384	12	hist_8_green	0.013445
3	hist_8_red	0.136617	13	entropy_green	0.010206
4	hist_6_green	0.124532	14	homogeneity_red	0.008832
5	hist_7_red	0.118374	15	hist_3_green	0.007654
6	hist_4_green	0.072164	16	hist_4_red	0.006617
7	hist_6_red	0.056674	17	hist_9_red	0.004262
8	hist_5_red	0.032635	18	homogeneity_gray	0.004088
9	hist_8_blue	0.022049	19	entropy_gray	0.003720
10	homogeneity_green	0.018141	20	dissimilarity_gray	0.003469

4.5.2 SHAP values

The **global SHAP** summary plot provides an overview of feature importances across all samples in the test dataset, where each point represents a SHAP value for an instance. The corresponding SHAP value is on the x-axis, representing the positive or negative impact on the model's prediction. On the y-axis, 20 features are sorted by decreasing order of importance depending on the sum of the absolute SHAP values across all instances. High values of the feature are represented in red, whereas low values are shown in blue.

The Global SHAP to understand feature importance across the entire dataset can be found in Figure 23. This graph shows the impact of each feature on the model output across all instances in the dataset. It can be observed that histogram from green and red components of RGB images are the features with higher SHAP value, indicating that, on average, these are the features that contribute more significantly to predictions. The features with lower or near-zero values are the ones with less impact.

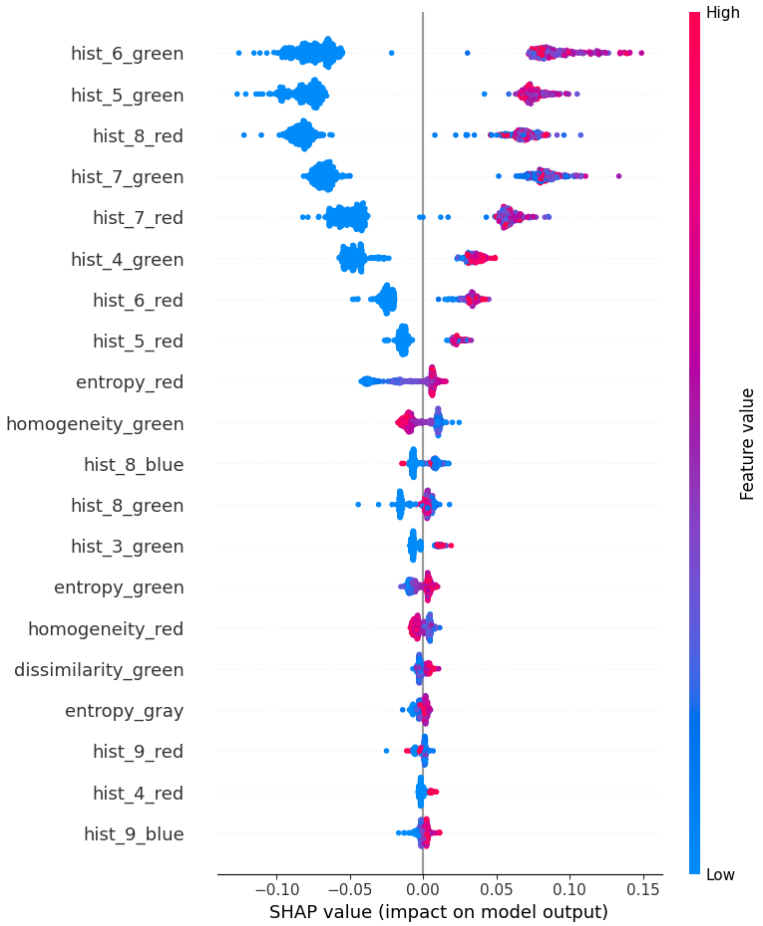


Figure 23. Summary plot of SHAP

Additionally, to compare SHAP values for uninfected and infected classes, SHAP summary plots separately for each class were obtained and are illustrated in Figure 24. This comparison helps to see how features contribute differently to each class's predictions and potentially highlight features that differentiate between classes.

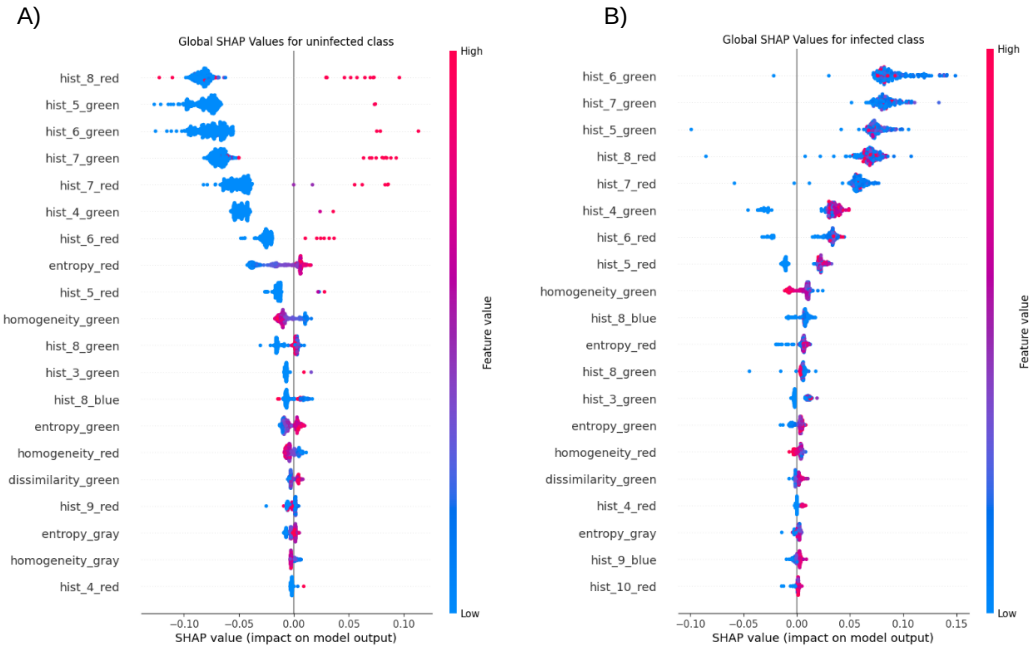


Figure 24. Global SHAP summary plots for A) uninfected class and B) infected class.

Features with larger absolute SHAP values for each class contribute more to predictions for that class compared to features with smaller absolute values. Apart from the magnitude of the value, it is interesting to evaluate the sign of the value. This SHAP values for a given feature can be positive or negative depending on how they influence the model's output.

In the uninfected class, high values of intensity in histograms of red and green images make the model less likely to predict an instance as uninfected, as SHAP values are negative. In contrast, for the infected class the same features have positive SHAP values, suggesting that higher values of intensity in histograms of red and green images tend to increase the model's prediction probability of the infected class. In other words, higher values of intensity in histograms of red and green images are associated with a decreased likelihood of being uninfected and an increased likelihood of being infected. Negative values for uninfected and positive values for infected indicate contrasting impacts on the model's predictions, highlighting the discriminatory power of these features in distinguishing between classes.

Apart from global SHAP, **local SHAP** force plots of 5 representative instances of both the uninfected and infected classes were visualized and are plotted in Figure 25 and Figure 26, respectively.

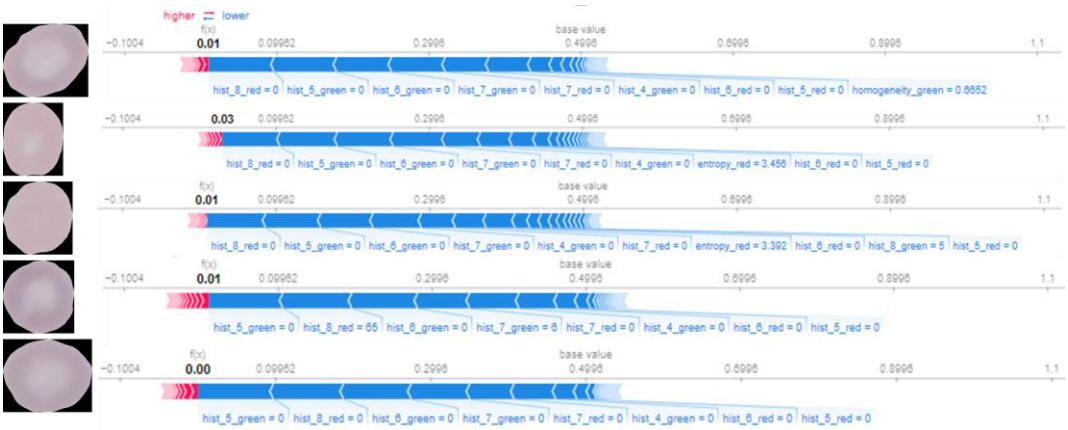


Figure 25. Local SHAP of 5 instances from uninfected class.

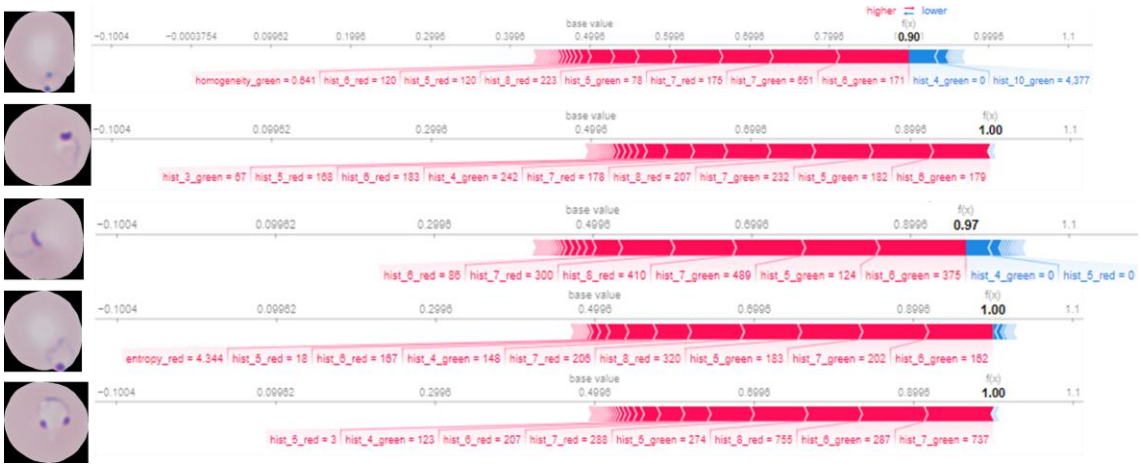


Figure 26. Local SHAP of 5 instances from infected class.

In the local SHAP force plots, the base value or expected value indicate the average model output over the training dataset, where the prediction would start without any influence from the features. Then, the prediction of the instance is pushed away from the base value by each feature depending on their SHAP value. The arrows show the contribution of each feature to the final prediction and the colour of the arrows indicate if the predictions are higher and are pushed towards infected class (red colour) or are pushed it lower towards uninfected class (blue colour). The final prediction of the model for the instance is showed by the end of the arrows.

By comparing these force plots, it can be observed how different features influence the prediction for each class, understanding the decision-making process of the model and identifying key features that differentiate the classes. Again, comparing local SHAP force plots of five instances of both classes, histograms from red and green components were identified as the most significant features to push the predictions towards uninfected and infected classes.

4.5.3 Interpretation of feature importance

Interpretation of importance of features in a Random Forest classification model with both feature importances and SHAP values led to a clear conclusion: histograms of red and green channels from RGB images capture and represent essential visual characteristics of the images to determine the presence of Plasmodium parasites.

A histogram of an RGB image represents the distribution of pixel intensities across its red, green, and blue colour channels. The extracted histogram from RGB image include statistics such as mean, variance, skewness, and kurtosis for each colour channel. Thus, the histogram feature provides a summary of the colour distribution in the image, reflecting its colour composition and variations in intensity.

Colour features, particularly the histogram features from red and green components of RGB images, are of significant importance to highlight visual cues that pathologists use to distinguish between uninfected and infected cells, allowing to obtain a reliable diagnose for the disease.

The effectiveness of colour features in differentiating between infected and uninfected images can be attributed to the fact that parasites turn a violet colour during staining. Specifically, the red and green components are important, but not the blue, as when the three components of RGB images are separated, the parasite is only observable in the red and green channels, as shown in the example in Figure 20.

While texture features are commonly used to describe spatial patterns of colour intensity (3), their importance in this study with a Random Forest model is not as pronounced as

that of colour features. However, entropy, homogeneity and dissimilarity are among the most significant texture features.

Regarding the geometric features, no substantial differences in the shape of the entire RBCs were expected between infected and uninfected cells, as the shape and size of the cells were not typically affected by the parasite infection. Generally, morphologic features are not used to distinguish infected from uninfected cells but are applied to determine parasite species and life stage (11).

4.6 Object detection model

To evaluate the object detection model to automatically draw the bounding boxes around RBCs, some images are shown in Figure 27.

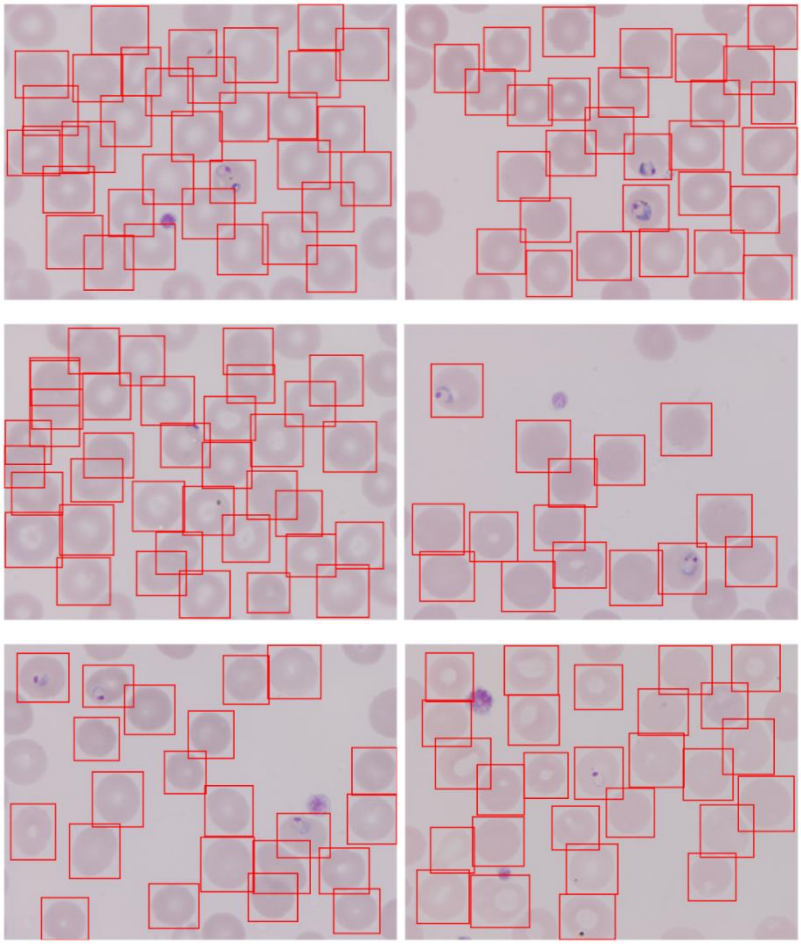


Figure 27. Evaluation images of the object detection model.

The object detection model provides good performance, as it can accurately detect and localize entire RBCs in images, without false positives (incorrectly detected objects, as other types of cells different from RBC), and without false negatives (entire RBCs that are missed).

Object detection of cells in microscopy images of malaria blood samples presents special challenges, as they have variations in illumination from the microscope, in cell shape, density, and colour from differences in sample preparation, as well as objects of uncertain class even for experts (15). Thus, to check robustness of the model it could be interesting to use new images with a variety of conditions, such as more patients' samples, different lighting, or complex backgrounds, as images densely populated with cells, which can result in confusing overlapping of boxes (32).

5 Conclusions

The manual annotation process requires significant human effort and is time-consuming. However, it is sometimes indispensable as it is essential for effectively training and validating models, achieving reliable results in automating the analysis of blood samples obtained from microscopes to diagnose malaria.

The segmentation of RBCs using the novel Segment Anything Model on malaria images was tested, yielding remarkable results. By selecting the mask with the highest score using the multimask option, a segmentation efficiency of 98.3% was achieved. Accurate extraction of cells is crucial for the subsequent classification model.

Several factors have been identified as causes of improper segmentation, including the location of the parasite at the periphery of the cell, cells with non-uniform colour (e.g., a whiter centre), and the proximity of cells to each other in images with high cell density.

A total of 80 colour, texture, and geometric features were extracted to train a machine learning model using a Random Forest algorithm after an undersampling process to address the highly imbalanced dataset. The model demonstrated excellent performance in classifying images into uninfected and infected classes, achieving 99.5% accuracy and a 99.2% F1 score.

Regarding feature importance, colour features were significantly more important than geometric and texture features in the Random Forest model's classification of infected and uninfected RBCs. Specifically, the histogram features from red and green components of RGB images are the most important features to differentiate both types of RBCs, as when the three components of RGB images are separated, the parasite is only observable in the red and green channels but not in the blue channel. The low importance of geometric features was expected, as the shape and size of infected and uninfected cells are generally not affected by parasite infection.

Finally, to completely automatise the diagnose, a model object detection was trained with the manual annotations previously obtained and results showed good performance.

Therefore, all the objectives set were carried out with satisfactory results. The sample analysis follows a three-step pipeline: detection, segmentation and classification. This multi-step approach would ensure that each cell is identified, its boundaries delineated and its characteristics correctly classified, obtaining an accurate malaria diagnosis by artificial intelligence. Furthermore, each step can be optimized independently, ensuring better overall performance.

Regarding the retrospective planning, only one notable deviation was made. Although it was not planned at the beginning of the project, the additional training of an object detection model was included to automate the cell detection process.

This work exemplifies how digital imaging technology, coupled with AI algorithms, is a promising tool for the diagnosis management of this important global health. It reduces the need for manual microscopy and improves the diagnostic accuracy of malaria, having a notable positive impact on the Sustainable Development Goals (SDGs) of Industry, Innovation, and Infrastructure, as well as Good Health and Well-being.

5.1 Future work

As future possibilities, it would be interesting to obtain patient samples' images from various laboratories, with different imaging equipments and sample preparation protocols. This could affect the visual appearance of images and, as consequence, the final performance of the model.

Although segmentation with SAM showed good performance, it would be necessary to properly segment the totality of cells. Morphological operations could be applied to improve the quality of binary masks. These techniques could remove small objects, fill small holes, detaching connected objects, etc., and enhance segmentation results.

Another interesting thing to explore would be the use of a model based on morphological characteristics to describe parasites inside the cells to know the specie identification of the parasite, as different species of parasites require different treatments. For example, *P. falciparum* is often more severe, which can lead to complications like cerebral malaria, severe anaemia and multi-organ failure. Also, this specie can be resistant to some antimalarial drugs. *P. vivax* and *P. ovale* form hypnozoites, a dormant phase in the liver that can cause relapses. Rapid identification of the specie is critical for immediate and specific treatment.

Likewise, identifying the stage of the parasite in the patient (e.g., ring, trophozoite, schizont or gametocyte), can provide information about the severity of the infection and the patient's prognosis, as late-stage parasites in the blood are associated with severe forms of malaria. The stage is also important for the treatment, as different antimalarial drugs target different stages of the parasite and antimalarial drug resistance can vary depending on the stage of the parasite.

Additionally, fully automating the detection and implementing a Graphical User Interface (GUI) would significantly enhance usability for healthcare professionals and avoid the need to understand coding practices or to be heavily trained in microscopic analysis.

6 Glossary

annotation: label parts of an image to categorize different interesting elements within the image. This is often used in machine learning to train algorithms.

bounding boxes: rectangular boxes drawn around objects in an image to identify them. They are a common method for image annotation.

classical machine learning: set of algorithms used to make predictions in data which are categorized into supervised learning and unsupervised learning.

GPU: Graphic Processing Unit. A piece of hardware in a computer designed for parallel processing. They are optimal for training machine learning models, as they can handle the large-scale matrix operations required.

image features: distinctive characteristics extracted from images that capture important information about their content. They are used in image classification tasks to feed into machine learning classifiers to learn discriminative patterns for different classes. Some commonly used types of image features are colour features, texture features and geometric features.

imbalanced dataset: dataset where the distribution of classes is heavily skewed, with one class more frequent than others.

mask: binary image that extracts specific regions of images. Each pixel is assigned either a value of 0 (black) for the background or 1 (white) for the region of interest.

plasmodium parasite: microorganism which cause malaria through the bite of infected female Anopheles mosquitoes. The most common species that can infect humans are *P. falciparum, P. vivax, P. ovale, P. malariae* and *P. knowlesi.* It has a complex life cycle with multiple stages that alternates between the mosquito vector and the human host.

peripheral blood smear: laboratory test that consist of spreading a drop of a patient blood onto a slide, staining it and examine it under a microscope.

random forest: supervised learning algorithm used in machine learning for classification and regression tasks based on decision trees. During the training phase, the algorithm learns from a dataset, which consists of features and their corresponding class labels.

red blood cells (RBCs): also known as erythrocytes, are the most abundant type of cells in the blood. In malaria patients, they are infected by parasites.

SHAP values: From SHapley Additive exPlanations, are a method based on cooperative game theory to explain the output of any machine learning. They help understand how each feature influences the machine learning model's output and identify relevant features.

segmentation: process of dividing an image into multiple regions that share similar characteristics, allowing the isolation of objects of interest from the background in an image.

supervised learning: type of machine learning where an algorithm is trained on a labelled dataset to predict the output for new data. For classification, the output variable is a category, whereas for regression, the output variable is a continuous value.

undersampling: resampling technique than consist of decreasing the number of instances in the majority class in an imbalanced dataset by randomly removing samples.

7 Bibliography

1. World Health Organization. World malaria report 2022. 2023.
2. Okumu F, Gyapong M, Casamitjana N, Castro MC, Itoe MA, Okonofua F, et al. What Africa can do to accelerate and sustain progress against malaria. PLOS Global Public Health. 2022 Jun 24;2(6):e0000262.
3. Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. Vol. 194, Translational Research. Mosby Inc.; 2018. p. 36–55.
4. Ikerionwu C, Ugwuishiwu C, Okpala I, James I, Okoronkwo M, Nnadi C, et al. Application of machine and deep learning algorithms in optical microscopic detection of Plasmodium: A malaria diagnostic tool for the future. Vol. 40, Photodiagnosis and Photodynamic Therapy. Elsevier B.V.; 2022.
5. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. 2023 Apr 5; Available from: http://arxiv.org/abs/2304.02643
6. Takian A, Raoofi A, Haghighi H. COVID-19 pandemic: The fears and hopes for SDG 3, with focus on prevention and control of noncommunicable diseases (SDG 3.4) and universal health coverage (SDG 3.8). In: COVID-19 and the Sustainable Development Goals. Elsevier; 2022. p. 211–34.
7. Delgado-Ortet M, Molina A, Alférez S, Rodellar J, Merino A. A deep learning approach for segmentation of red blood cell images and malaria detection. Entropy. 2020 Jun 1;22(6):1–16.
8. Rosado L, Da Costa JMC, Elias D, Cardoso JS. Automated Detection of Malaria Parasites on Thick Blood Smears via Mobile Devices. In: Procedia Computer Science. Elsevier B.V.; 2016. p. 138–44.
9. Shambhu S, Koundal D, Das P, Hoang VT, Tran-Trung K, Turabieh H. Computational Methods for Automated Analysis of Malaria Parasite Using Blood Smear Images: Recent Advances. Comput Intell Neurosci. 2022;2022.
10. Molina A, Rodellar J, Boldú L, Acevedo A, Alférez S, Merino A. Automatic identification of malaria and other red blood cell inclusions using convolutional neural networks. Comput Biol Med. 2021 Sep 1;136.
11. Nugroho AS, Winarta T, Wibisono Y, Galinium M, Rozi IE, Asih PBS. Morpho-geometrical feature extraction of thin blood smear microphotograph for malaria plasmodia species and life stage determination. In: 2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020. Institute of Electrical and Electronics Engineers Inc.; 2020. p. 95–100.
12. Sumi AS, Nugroho HA, Hartanto R. A Systematic Review on Automatic Detection of Plasmodium Parasite. International Journal of Engineering and Technology Innovation. 2021;11(2):103–21.
13. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018 Nov 1;34(21):3711–8.
14. Rajab S, Nakatumba-Nabende J, Marvin G. Interpretable Machine Learning Models for Predicting Malaria. In: 2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2023. Institute of Electrical and Electronics Engineers Inc.; 2023.
15. Hung J, Carpenter A. Applying Faster R-CNN for Object Detection on Malaria Images. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society; 2017.

16. Maturana CR, de Oliveira AD, Nadal S, Bilalli B, Serrat FZ, Soley ME, et al. Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. Vol. 13, Frontiers in Microbiology. Frontiers Media S.A.; 2022.

17. Jan Z, Khan A, Sajjad M, Muhammad K, Rho S, Mehmood I. A review on automated diagnosis of malaria parasite in microscopic blood smears images. Multimed Tools Appl. 2018 Apr 1;77(8):9801–26.

18. Uzun Ozsahin D, Mustapha MT, Bartholomew Duwa B, Ozsahin I. Evaluating the Performance of Deep Learning Frameworks for Malaria Parasite Detection Using Microscopic Images of Peripheral Blood Smears. Diagnostics. 2022 Nov 1;12(11).

19. Ceusters W, Smith B. Malaria diagnosis and the Plasmodium life cycle: the BFO perspective. Nature Precedings. 2010 Jan 6:1-.

20. Lowelp BS, Jeffa NK, New L, Pedersen3 C, Engbaek3 K, Marsh K. Acridine orange fluorescence techniques as alternatives to traditional Giemsa staining for the diagnosis of malaria in developing countries. Vol. 90. 1996.

21. Lazrek Y, Florimond C, Volney B, Discours M, Mosnier E, Houzé S, et al. Molecular detection of human Plasmodium species using a multiplex real time PCR. Sci Rep. 2023 Dec 1;13(1).

22. Tedla M. A focus on improving molecular diagnostic approaches to malaria control and elimination in low transmission settings: Review. Vol. 6, Parasite Epidemiology and Control. Elsevier Ltd; 2019.

23. Sukumarran D, Hasikin K, Mohd Khairuddin AS, Ngui R, Wan Sulaiman WY, Vythilingam I, et al. Machine and deep learning methods in identifying malaria through microscopic blood smear: A systematic review. Vol. 133, Engineering Applications of Artificial Intelligence. Elsevier Ltd; 2024.

24. White M, Marais P. Supervised learning and image processing for efficient malaria detection. Cape Town; 2019.

25. Alnussairi MHD, İbrahim AA. Malaria parasite detection using deep learning algorithms based on (CNNs) technique. Computers and Electrical Engineering. 2022 Oct 1;103.

26. Hemachandran K, Alasiry A, Marzougui M, Ganie SM, Pise AA, Alouane MTH, et al. Performance Analysis of Deep Learning Algorithms in Diagnosis of Malaria Disease. Diagnostics. 2023 Feb 1;13(3).

27. Acherar A, Tantaoui I, Thellier M, Lampros A, Piarroux R, Tannier X. Real-life evaluation of deep learning models trained on two datasets for Plasmodium falciparum detection with thin blood smear images at 500x magnification. Inform Med Unlocked. 2022 Jan 1;35.

28. Yang F, Poostchi M, Yu H, Zhou Z, Silamut K, Yu J, et al. Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears. IEEE J Biomed Health Inform. 2020 May 1;24(5):1427–38.

29. Maturana CR, de Oliveira AD, Nadal S, Serrat FZ, Sulleiro E, Ruiz E, et al. iMAGING: a novel automated system for malaria diagnosis by using artificial intelligence tools and a universal low-cost robotized microscope. Front Microbiol. 2023;14.

30. Tobias RR, Carlo De Jesus L, Mital ME, Lauguico S, Guillermo M, Vicerra RR, et al. Faster R-CNN Model with Momentum Optimizer for RBC and WBC Variants Classification. In: LifeTech 2020 - 2020 IEEE 2nd Global Conference on Life Sciences and Technologies. Institute of Electrical and Electronics Engineers Inc.; 2020. p. 235–9.

31. Basu D, Dastidar SG. Molecular Dynamics and Machine Learning reveal distinguishing mechanisms of Competitive Ligands to perturb α,β-Tubulin. Comput Biol Chem. 2024 Feb 1;108.

32. Loh DR, Yong WX, Yapeter J, Subburaj K, Chandramohanadas R. A deep learning approach to the screening of malaria infection: Automated and rapid cell counting, object detection and instance segmentation using Mask R-CNN. Computerized Medical Imaging and Graphics. 2021 Mar 1;88.

8 Appendix

8.1 Segmentation efficiency

Good performance in segmentation of RBCs was obtained with SAM. Nevertheless, some images were discarded for various reasons. On the one hand, 12 images out of a total of 8.750 were eliminated as their incorrect segmentation resulted in a mistaken label. The discarded images are shown in Figure S1. Typically, these images shared the characteristic of the parasite being in the outermost part of the cell, with two cases where the parasite belonged to an adjacent cell. Despite this, 99.9% of the images were segmented accurately without affecting the class labels.

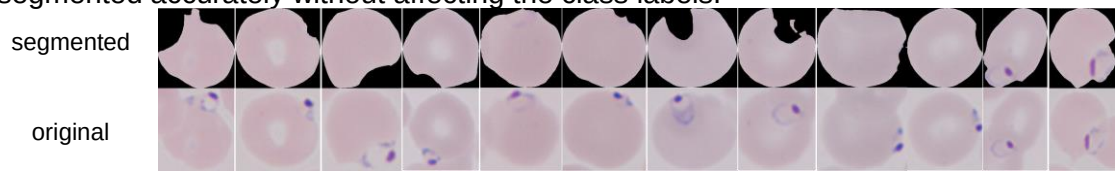


Figure S1. Discarded images after segmentation due to due to changes in class labels.

On the other hand, a total of 137 images were also discarded due to incorrect masks. In general, these images are characterized by belonging to areas with high cell density or non-uniform cell coloration. Examples of discarded images are presented in Figure S2.

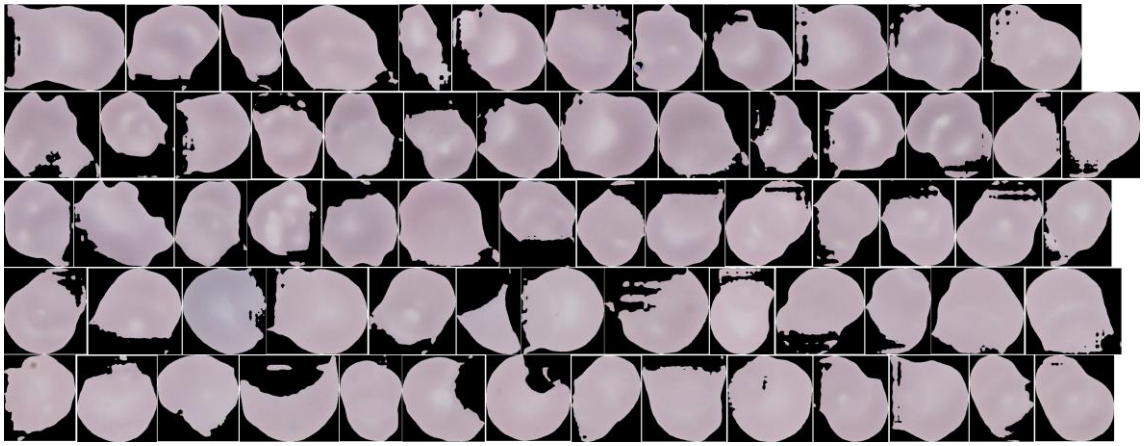


Figure S2. Some examples of discarded images after segmentation due to incorrect masks.

8.2 Binary images optimization

Binary images are needed to calculate geometrical features such as area, perimeter, circularity and eccentricity. In order to obtain these binary images from grayscale images, thresholding was applied. Therefore, selecting an appropriate threshold value is essential to separate infected and uninfected RBCs from the background but this value could vary depending on factors such as light and cell staining.

To obtain fully binarized RBCs, the threshold value was optimized in infected RBCs by visually checking binary images generated when applying threshold values of 128, 90, 80, 70 and 60. The value of 128 is commonly used as starting point because it represents the mid-range intensity value for an 8-bit grayscale image, where pixel intensities range from 0 (black) to 255 (white). Results presented in Figure S3 indicate that the optimal threshold value to binarize infected RBC images was 60, as in other cases, the parasite was considered as background due to its dark colour.

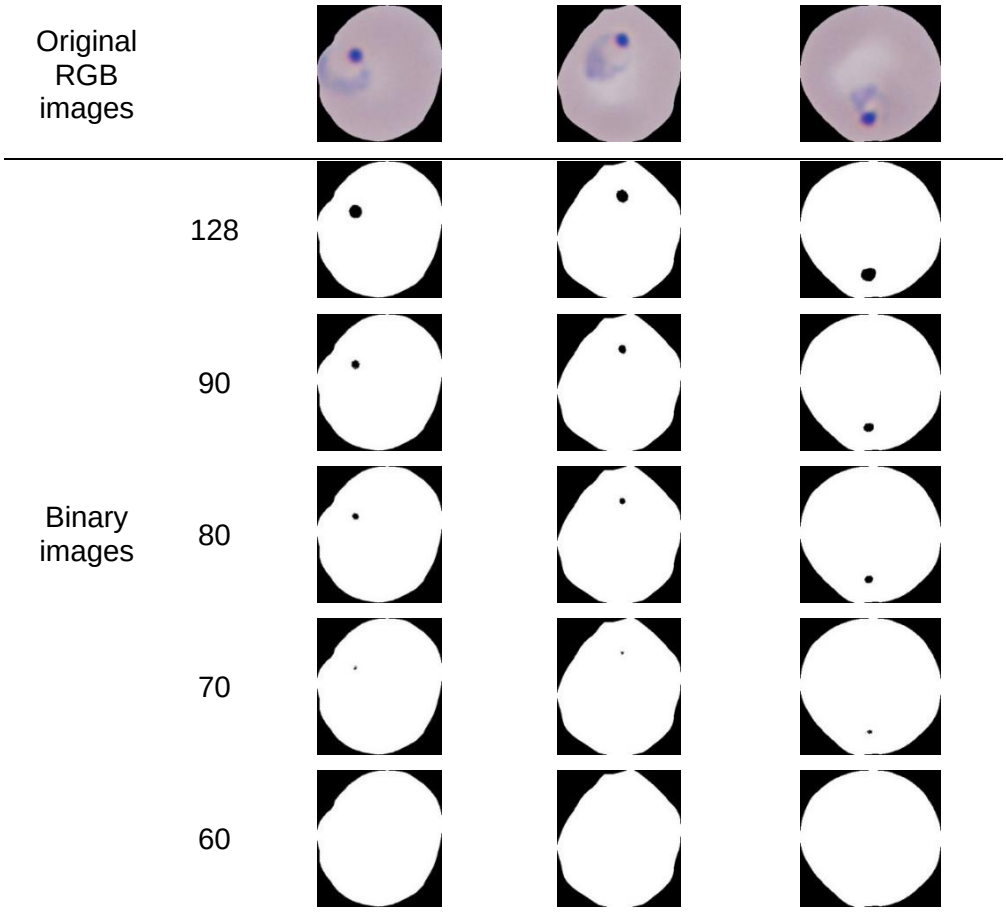


Figure S3. Optimization of the threshold value to obtain binary images. Values of 128, 90, 80, 70 and 60 were tested.

8.3 Missclassified images in the classification model

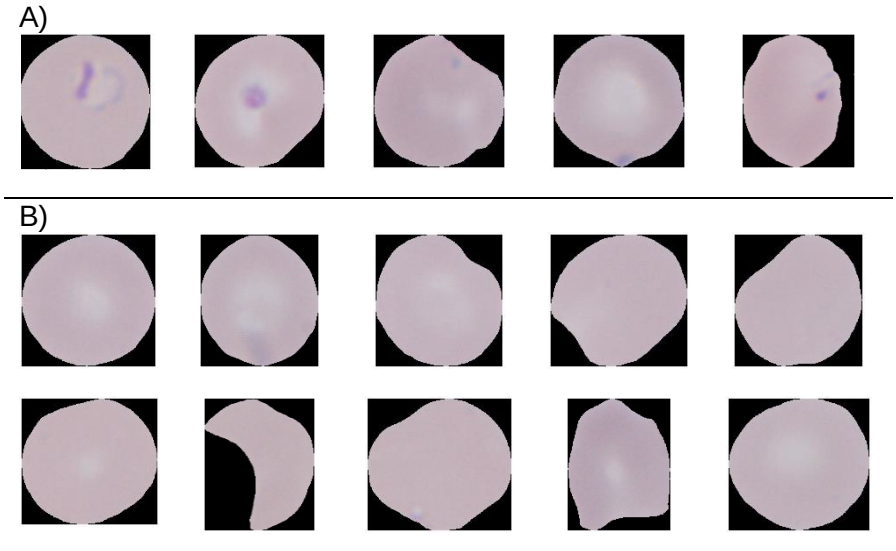


Figure S4. Missclassified images obtained in the Random Forest classification model. A) False negatives, B) False positives