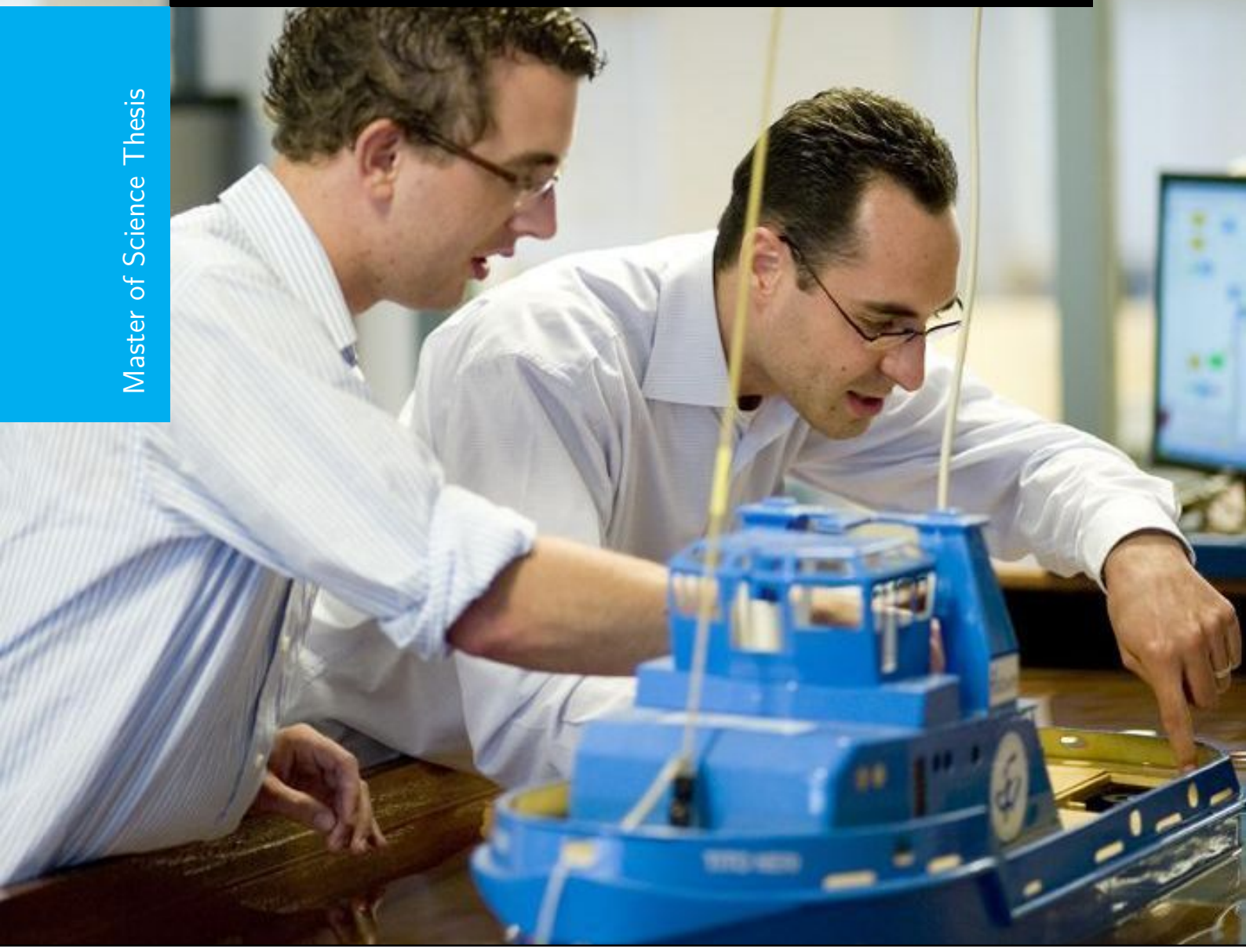
Improving in-the-field performance of automated malaria diagnosis

M. Pors

Master of Science Thesis



TU Delft Delft University of Technology

Delft Center for Systems and Control

Improving in-the-field performance of automated malaria diagnosis

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

M. Pors

August 13, 2024

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Abstract

Malaria remains a leading cause of mortality, particularly in regions with limited access to healthcare. Despite the availability of diagnostic methods such as microscopy, Rapid Diagnostic Tests (RDT), and Polymerase Chain Reaction (PCR), these methods face challenges in accuracy and practicality, especially in field conditions. This thesis investigates the use of automated, Computer-aided diagnostics (CAD) to improve in-the-field malaria diagnosis. While CAD shows promise, significant challenges remain in dealing with image imperfections and data variability. The study focuses on the use of blob detectors and Zernike decomposition to improve the precision of the CAD system. Additionally, data quality is investigated to evaluate the reliability of malaria diagnostic pipeline. The results suggest that Zernike decomposition is effective in identifying shapes most likely to correspond to parasites, enhancing the blob detector's precision. Furthermore, the use of data stratification is shown to reduce the variance between models during training, which indicates that it can generalise better to unseen data. This work contributes to the ongoing effort to develop a robust, field-deploy-able malaria diagnostic tool by incorporating prior knowledge of the malaria parasite its shape to improve precision and ensuring reliability by evaluating data quality.

Table of Contents

1 Introduction **1**

1-1 Roadmap . . . 3

2 Recognition of Malaria **5**

2-1 Manual diagnosis using brightfield microscopy . . . 5

2-2 Staining of blood samples . . . 6

2-3 Examination of blood sample . . . 6

2-4 Malaria parasite characteristics . . . 7

2-5 Performance evaluation of technicians . . . 7

2-6 Conclusion . . . 8

3 Automated diagnosis **9**

3-1 Convolutional Neural Networks (CNN's) . . . 9

3-2 Feature Detection . . . 11

3-2-1 Feature detectors . . . 12

3-2-2 Scale space . . . 13

3-2-3 Zernike Decomposition . . . 13

3-3 Conclusion . . . 14

4 Dealing with errors and imperfections **15**

4-1 Image quality . . . 15

4-1-1 Noise . . . 15

4-1-2 Low image contrast . . . 15

4-1-3 Colour variation . . . 17

4-1-4 Artefacts . . . 18

4-2 Data quality . . . 18

4-2-1 Data size . . . 19

4-2-2 Data annotations . . . 19
4-3 Performance metrics . . . 19
4-3-1 Binary classification . . . 19
4-3-2 Area Under Curve (AUC) and Area Under Precision-Recall Curve (AUPRC) 20
4-3-3 Metrics based on counts . . . 20
4-4 Conclusion . . . 21

5 Methods **23**
5-1 Preliminaries . . . 23
5-1-1 Malaria parasite . . . 23
5-1-2 Terminology . . . 24
5-1-3 Benchmark overview . . . 25
5-2 Roadmap . . . 25
5-3 Data Partitioning . . . 27
5-4 Data Exploration . . . 28
5-4-1 Parasite Density . . . 28
5-5 Adapted preselection . . . 29
5-5-1 Zernike Decomposition . . . 29
5-5-2 Feature detector . . . 31
5-6 Classification . . . 34
5-6-1 Data Stratification . . . 34
5-7 Evaluation . . . 35
5-7-1 Preselection performance . . . 35
5-7-2 Classification performance . . . 35
5-7-3 In-the-field performance . . . 37
5-8 Summary . . . 37

6 Results **39**
6-1 Improved preselection precision . . . 39
6-2 Comparison of in-the-field classification performance . . . 41
6-3 Effect of data stratification on classification performance . . . 41
6-4 Identification of parasite shapes using Zernike Decomposition . . . 42
6-5 Conclusion . . . 43

7 Conclusion **45**

Bibliography **47**

Glossary **55**
List of Acronyms . . . 55
List of Symbols . . . 55

Chapter 1

Introduction

Malaria is one of the leading causes of death in South Africa, attributed to poor access to healthcare, understaffing of public health facilities, and lack of equipment [49, 74, 81]. The WHO African region accounts for approximately 95% of the total 247 million malaria infections [74], of which less than 29% receives proper treatment [81]. Despite the WHO its efforts, the mortality rate (deaths per 100 000 people at risk) did not decrease significantly (1.3%) as of 2015 [74]. Therefore, it remains vital to keep developing (new) diagnostic methods that aid in the eradication of malaria.

Rapid and accurate diagnosis can be achieved with current diagnostic methods, yet their performance deteriorates when applied in-the-field. The predominantly used methods include: microscopy, which involves manual inspection of blood samples for malaria parasites; Rapid Diagnostic Tests (RDT), which involve applying a drop of blood to a paper strip that shows a visible line if the patient is infected; and Polymerase Chain Reaction (PCR), which effectively creates countless copies of a DNA sample to allow for careful analysis. Despite microscopy being the golden standard, it is labour-intensive, and its efficacy depends on factors such as the quality of the blood sample, available equipment, and skill level of the handler [23, 34]. Low skill levels due to inadequate training can result in misdiagnosis; studies have reported that 36% to 41% of the cases are misdiagnosed [49, 81]. Moreover, microscopy is usually done in a laboratory, which impedes diagnostics where resources are low [34]. RDT are widely used as an alternative if microscopy diagnosis cannot be provided, due to their rapid deployment and quick results. However, the performance is severely affected by climate conditions and it only detects malaria if the density is > 100 parasites/μl, compared to > 5 parasites/μl for manual microscopy [71]. PCR, however, can detect malaria for densities as low as 0.004 parasites/μl, but the use is limited due to its high cost and storage conditions [49]. Hence, whereas microscopy its performance is affected by humans, the alternatives, RDT and PCR, are affected by external factors. This presents the possibility to improve upon microscopy diagnostics by either improving the performance of the handler, or by reducing the dependence on trained personnel.

Automated diagnosis has been proposed to reduce the dependence on trained personnel, referred to as Computer-aided diagnostics (CAD), which reduces the amount of labour and

can lead to more reliable diagnostic results [7, 41, 50]. CAD consists of four main steps: pre-processing (e.g. filtering out noise), feature extraction (retrieving relevant information from an image), classification (using the provided information to evaluate whether this is a parasite or not), and performance evaluation (verification based on ground truth) [41]. Numerous CAD implementations have been proposed [41, 50] to achieve accurate diagnosis ($> 96\%$ [60]), typically by implementing increasingly complex machine learning and deep learning algorithms on high-end devices. To mitigate the dependence on high-end devices, algorithms are proposed that can be implemented on smartphones [79]. Using readily available technology such as smartphones facilitates in-the-field deployment [55], yet this gives rise to a new challenge: processing of in-the-field data.

Data acquired in-the-field gives rise to complications such as poor image quality, available computing power, and variations in the images due to preparation or contamination of the sample [23, 35, 37, 78]. To account for these imperfections, the traditional approach in CAD is to implement a series of manually tuned filters to acquire a dataset that is as homogeneous as possible. However, the used datasets are acquired in different locations and conditions, introducing deviations between datasets due to differences in sample preparation and imaging settings [50, 53]; these variations deteriorate the homogeneity of the datasets. Since CAD methods proposed in the literature are typically evaluated on one dataset only, it is not clear if these methods generalise to other datasets as well.

To mitigate the need for manual tuning and improve the generalisability of the automated diagnostic algorithm, focus has shifted from the traditional approach to neural networks by replacing the feature extraction step or providing an end-to-end solution [41]. Neural networks typically outperform traditional approaches [60], although the performance deteriorates when dealing with unseen data [78]. This hinders application in-the-field since datasets acquired in this setting contain more imperfections (unseen data) than datasets acquired in a clinical setting [34,59]. Although algorithms can be trained using in-the-field data [38], this is impeded by the scarcity of publicly available in-the-field data [78]. To deal with these imperfections, or artefacts, several approaches have been proposed, such as: manually removing images with imperfections [34], using fluorescent imaging [33], or setting a threshold for the minimum number of detected parasites to account for some of the detections being artefacts [78]. Since the goal is to reduce manual labour, and the fact that in-the-field labs are typically only equipped with brightfield microscopes, the first two methods are not sufficient. The third method does not require additional manual labour or equipment, yet requires tuning of the threshold parameter and assumes that the number of parasites will equivalent for all images in the dataset. Therefore, dealing with imperfections such as artefacts remains an open area of research.

It is clear that quality of images in a dataset affect the performance of CAD algorithms, yet the performance is also affected by the quality of the data itself. For example, the performance of Convolutional Neural Network (CNN)s depends strongly on the amount of data available [36], and if there is a class imbalance in the dataset, e.g., if there are more samples of negative images than of positive images, the performance can be affected as well [54]. In the latter case, it is also important to carefully select which performance metric is used as it can overestimate the performance [54, 56, 61].

In conclusion, CAD has shown to be a promising approach to automate the diagnosis of malaria, yet requires further research in order to successfully implement it in-the-field. First

off, it is unclear how to distinguish between parasites and imperfections such as artefacts. Second, data quality is not always taken into account when evaluating performance of a CAD algorithm. Lastly, when evaluating, metrics are used that give a biased representation of the performance. Therefore, this thesis aims to find characteristics, or features, that set imperfections apart from parasites to improve in-the-field performance, and explores other error sources such as data quality and performance metrics to evaluate how they affect the performance of automated diagnostic algorithm.

1-1 Roadmap

This thesis is structured into four main parts: background chapters, methods, results, and conclusions. The background chapters provide context and reviews relevant literature; the methods chapters discusses how the research is performed; the results chapter presents the outcomes of the research; and the conclusion summarises how the results have contributed to the field. The following chapters will together form the background part of this thesis. Chapter 2, will review how manual diagnosis is performed to analyse what characteristics are used to recognise malaria, and what guidelines the microscopist need to follow to ensure that the patient gets the correct treatment. For example, it is essential to know the severity of the infection for proper treatment, therefore it is also important to analyse how this is assessed. Next, it is reviewed how this process is ported to automated diagnosis in Chapter 3. Lastly, Chapter 4 reviews how past studies have dealt with image- and data quality and how they evaluated the performance.

Chapter 2

Recognition of Malaria

Characteristics of a malaria parasite such as size, shape and colour are vital when learning how to distinguish between parasites and other objects; this is true for both lab technicians and computer algorithms. However, whereas lab technicians intuitively know what to look for if a parasite is described as a *purple, ring-shaped object*, computer algorithms first need to learn to recognise characteristics like colour and shape. These characteristics are called *features*. This chapter reviews the manual diagnosis process to better understand which characteristics are decisive in recognising malaria, and what procedures are in place to establish a diagnosis.

2-1 Manual diagnosis using brightfield microscopy

Due to its availability in laboratories, versatility in diagnostics and simplicity in handling, brightfield microscopy remains the golden standard for manual malaria diagnosis [23, 73]. In brightfield microscopy the sample is typically illuminated from below and observed from above. As a result, objects in the sample appear darker than the rest of the sample due to the absorption and reflection of light. This allows technicians to spot objects that would otherwise remain undetected. The diagnosis through brightfield microscopy (roughly) consists of three important parts: sample preparation, handling the microscope, and reading of the sample, i.e., examining whether parasites are present in the sample.

The types of samples can range from stool samples to urine samples to blood samples, of which the latter is a standard option for diagnosing malaria. The blood smear consists of a single layer or multiple layers of blood, called thin and thick blood smears, respectively. The fact that more blood is used makes the thick blood smear more sensitive for parasite detection and quantification [71]. However, the image will inherently contain some defocused objects as a result of the multiple layers. In thin blood smears, red blood cells are preserved and therefore are more suitable to classify different species of parasites [73]. Technicians are therefore instructed to first examine a thick blood smear to derive if a patient is infected or not, and only examine a thin blood smear if the parasite species cannot be confirmed in the thick blood smear.

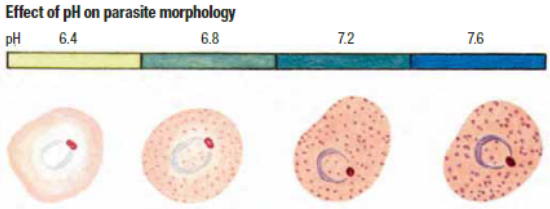


Figure 2-1: Deviations in colour balance due to staining. If staining is done correctly, the pH should be around 7.2 [71].

2-2 Staining of blood samples

To ensure that the technician is able to observe parasites in a blood smear under a microscope, both thin or thick, a dye is applied to the blood smear [73]. This dye stains each part of the parasite with a distinct, more intense colour, which enhances the contrast between the parasite and background [23] [58]; this process is called staining. Due to a lack of standardisation and quality control, the staining process differs between laboratories, resulting in variations in contrast and colour [35] [49]. In addition, artefacts can be introduced during the staining process, such as blurring, dirt contamination, or smudges due to incorrect application of the dye [37]. In Figure 2-1 it is show how the colour balance of a sample is affected by deviations in the staining process. Hence, although staining is highly recommended for malaria detection [73, 74], it is also a possible source of imperfections that can complicate diagnosis.

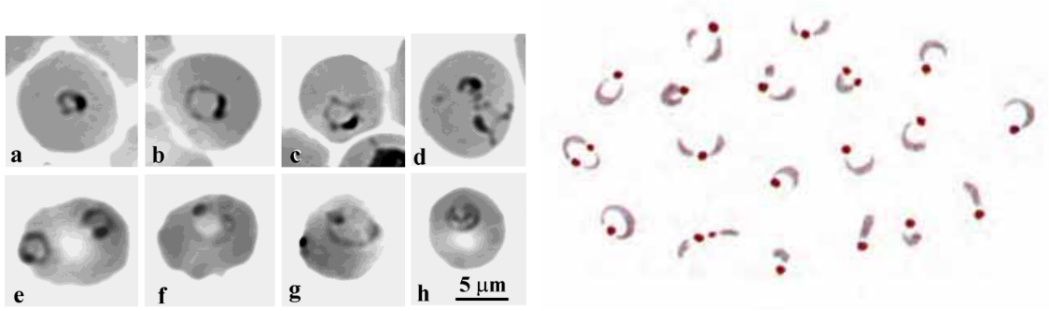
2-3 Examination of blood sample

Once the blood smear is prepared, it is placed under the microscope for examination. The microscope is typically fitted with multiple objectives so that the sample can be examined under different magnifications. It is recommended to use 10× oculars, but 7× oculars are used in practice as well to enable the technician to scan a larger area in one go. Starting with the thick blood smear, a lower magnification (usually 40×) is used to select a part of the blood smear that is free of contamination and is properly stained [71]. This part of the blood smear, also called the Region of Interest (ROI), is then examined under a higher magnification (usually 100×). To ensure a reliable diagnosis, the technician should examine at least 100 different areas within the ROI, referred to as *fields*, and confirm that at least $15 - 20$ White Blood Cell (WBC) are present in each field. The former is used to reduce false negative diagnoses for low parasite densities. Parasite density D, or parasitaemia, is defined as

$$\mathrm{D} = \frac{c_p}{V_{bs}} = \frac{c_p \times N_{wbc}}{c_{wbc}} = \frac{c_p \times 8000}{c_{wbc}}, \tag{2-1}$$

c_p denotes the count of parasites, N_{wbc} and c_{wbc} denote the total number and counted WBC's respectively. The latter is assumed to be 8000 if the actual count is not available [71]. Hence, for 100 images the minimum detectable parasitaemia, assuming an average of 20 WBC's, is [51]:

$$\mathrm{D}_{\min} = \frac{1 \times 8000}{100 \times c_{wbc}} = \frac{1 \times 8000}{100 \times 20} = 4. \tag{2-2}$$



(a) **(b)**

Figure 2-2: a) Red blood cells infected ring-stage malaria parasites. (a-g) display parasites at different points in their life cycle. (h) displays an earlier stage. adapted from [13] b) Illustration of ring-stage malaria parasite appearances in thick blood films. The red dots resemble the nucleus, the cytoplasm is attached to the nucleus. Adapted from [71]

The probability of a false negative (False Negative Rate (FNR)) is calculated as follows: [72]

$$\text{FNR} = \left(\frac{N-n}{N}\right)^{C_p}, \tag{2-3}$$

where C_p denotes the total count of parasites for the total number of images N and n denotes the number of samples taken from the population. Hence, after analysing 100 out of 1000 images

$$\text{FNR} = \left(\frac{1000-100}{1000}\right)^{10} = 34.8\%, \tag{2-4}$$

where $C_p = 10$ corresponds to the minimum detectable parasitaemia, D = 4.

2-4 Malaria parasite characteristics

After an ROI is selected, the technician scans the area to identify parasites. During the lifetime of a parasite it goes through several stages, where each stage of the parasite can be recognised through specific characteristics such as shape and size. In Figure 2-2b, the different parts of the parasite are shown, which can be used to recognise them. The first thing to look for is the *chromatin*, a part of the nucleus, which is always present in a parasite. If one or more nuclei are present, the next thing to check is their size and shape. The nucleus is usually round and the parasite's size ranges between $2 - 3.7\mu m$ [13], which can be seen in Figure 2-2a. Another common characteristic is the *cytoplasm*, which is attached to the nucleus and is typically ring shaped.

2-5 Performance evaluation of technicians

To guarantee that technicians' performance meets the established standards, the World Health Organisation (WHO) prescribes how to assess their performance. Malaria microscopists can have different levels of expertise, which are evaluated using performance metrics shown in

Table 2-1: Accreditation levels for malaria microscopists.

Accreditation level	Detection accuracy	Parasite quantitation (25% of true count)	FPR
Level 1 (Expert)	$\geq 90\%$	$\geq 50\%$	$\leq 2.5\%$
Level 2	$< 90\%$	$< 50\%$	$\leq 5\%$
Level 3	$< 80\%$	$< 40\%$	$\leq 10\%$
Level 4	$< 70\%$	$< 30\%$	$\leq 20\%$

FPR denotes False Positive Rate. Metrics are evaluated on image level. Adapted from [73].

Table 2-1. These metrics are similar for machine learning algorithms as will be discussed in Section 4-3, however, there are some subtle differences: First, detection accuracy (2-5a), is often misinterpreted as accuracy of classification; detection accuracy only takes into account if the microscopist correctly diagnosed a slide, as opposed to individual (potential) parasites within each image [72]. Second, precision (2-5c), and recall (2-5d)(sensitivity) are often given the same weight, although they are not necessarily equally important. As shown in Table 2-1, misdiagnosis of 10% of the slides is acceptable for an expert level microscopist, yet the false positive rate (2-5b) should be $\leq 2.5\%$; this infers that more weight is put on reducing false positives.

$$\text{accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2-5a}$$

$$\text{specificity} = \text{FPR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2-5b}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2-5c}$$

$$\text{recall} = \text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2-5d}$$

2-6 Conclusion

To detect malaria we first need to identify objects in the image, which can be found by searching for parts of the image where the intensity is low. The object is identified as a parasite if it matches the description of the malaria parasite, e.g. if it has a similar shape and colour. Note that the colour can be affected by human interference: improper staining will shift the colour balance of the sample. The shape, however, remains the same. Moreover, by using thick smears, the probability of detecting a parasite is improved due to the larger volume of the blood sample. Lastly, it can be concluded from that technicians and algorithms can be evaluated using the same performance metrics. The WHO uses these metrics as well and prescribes that minimising false positives outweighs minimising false negatives. Moreover, the quantification does not have to be spot on as long as the other performance requirements are met. Hence, the recommended approach for detecting malaria is by looking for similar shaped objects in thick blood smears, while only classifying them as parasites if it is evident to avoid false positives.

Chapter 3

Automated diagnosis

Automated diagnosis is proposed to reduce the dependence on trained personnel and make the diagnostic results more reliable through implementation of algorithms. The process of automated diagnosis is, too a certain extent, analogous to that of manual diagnosis: First, the image is scanned for object with similar characteristics as malaria parasites (Feature extraction). Second, the potential parasites are inspected more closely to label it as a parasite (Classification). Lastly, the predicted labels are compared to the true labels to evaluate the algorithms performance (Performance evaluation). Hence, the goal of automated diagnosis is to find and implement an algorithm, or algorithms, that can perform these tasks. Although these steps were implemented as separate steps by traditional approaches for automated diagnosis, following the seminal work of Krizhevsky et al. [39], the focus has shifted to deep learning approaches that deal with multiple, or all of these steps at once: Convolutional Neural Networks (CNN's) [82].

3-1 CNN's

A CNN trains a set of filters to detect features instead of using a separate feature extraction step. These features can be used as input for traditional classification approaches such as Support Vector Matrix (SVM), or a CNN can be designed to classify these features as well to provide an end-to-end solution [41, 82]. In Figure 3-1 an overview with examples are shown that visualise the workings of a typical CNN architecture used in image classification.

CNN's usually outperform traditional machine learning algorithms and are therefore widely applied in Computer-aided diagnostics (CAD) [4, 5, 12, 60], yet are typically considered to be computationally expensive [12, 36, 82] which might reduce its applicability in-the-field where resources are limited. Nevertheless, they have been implemented on low-end (compared to, e.g. laptops) devices such as Raspberry Pi's [53] and smartphones [78, 79, 81], which are readily available even in settings where resources are limited [55]. Even though the available computational resources on a smartphones are limited when compared to (super)computers, the reported accuracy ranges between 96.5% and 99.5% [7, 52, 79]. These algorithms are

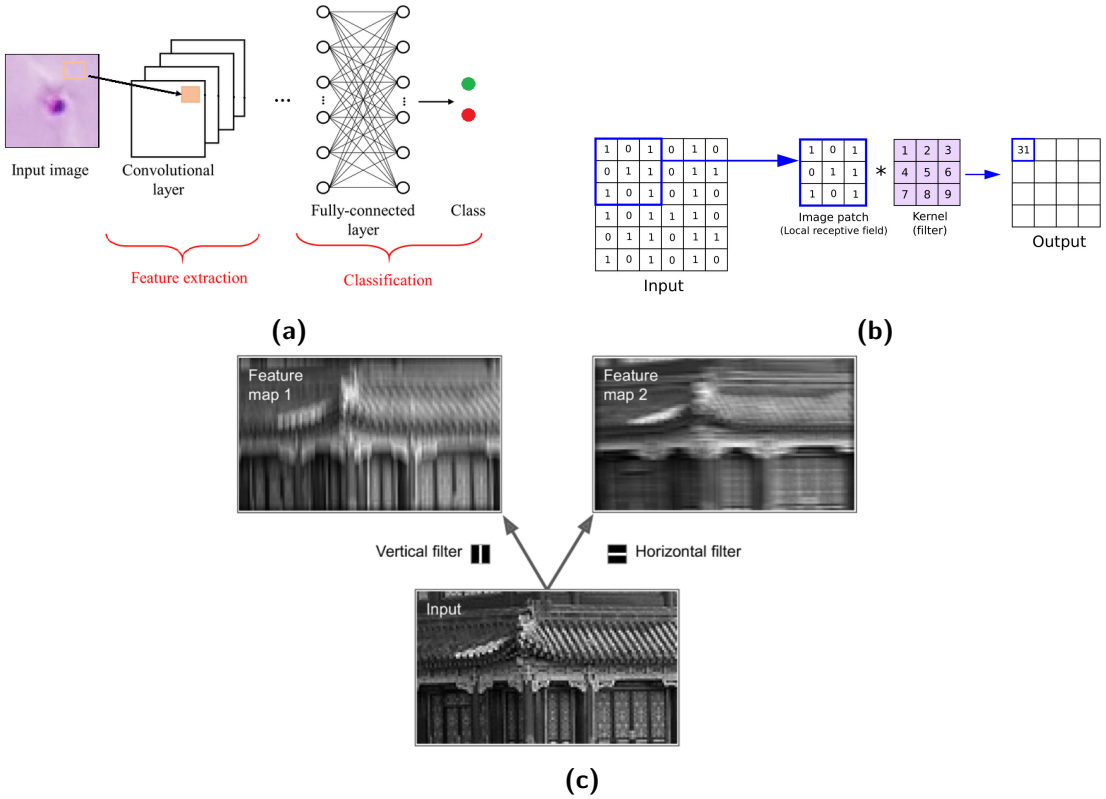


Figure 3-1: Overview of a typical CNN architecture and its workings. a) General structure of the CNN: the image is used as input for a convolutional layer consisting of multiple kernels, which output is again used for another convolutional or fully-connetected layer. b) An example of how an input is transformed to the output by the kernel of a convolutional layer, which is also visualised using a real world image in c). Adapted from [1, 26, 31].

trained on images that are acquired in a laboratory setting. The algorithm proposed in [79] is evaluated using in-the-field data as well, which showed a relatively low specificity of 51.1%, compared to 74% for lab data. This drop in performance is often attributed to parasite-like staining artefacts [34, 78]. Unseen data that is dissimilar from the training data can indeed deteriorate the performance [83], yet is not guaranteed to be the only source of error: error sources such as shift and rotation-invariance [40, 80] or data imbalance [54] are typically not ruled out. The lack of interpretability of CNN's due to its black-box design not only complicates finding the error sources, it makes it difficult to predict whether it can deal with imperfections that are not present in the training data, but are present in new, unseen data such as in-the-field data. Hence, CNN's in-the-field applicability is not necessarily reduced due limited computational resources, yet the lack of interpretability impedes finding a solution that generalises to other datasets.

Hybrid CNN approaches are proposed in [77–79] to reduce computational complexity by only selecting certain parts of the image based on the image intensity.Image patches where the image intensity is low are selected based on the fact that objects, including parasites, in brightfield images appear darker than the background. This filters out most of the background, yet it does not filter out other objects in an image, such as artefacts. The traditional approach to filter out irrelevant objects is to find *features* that are correlated to malaria parasites, and filter out objects that are not or weakly correlated to these features [21, 22]. Although previous studies have attempted to find features corresponding to malaria, [20–22], they did not focus on *why* these features would be relevant for detecting parasites and did not investigate if the features were invariant to image transformations such as rotation or scaling. In other medical fields, however, studies have been able to relate certain features to the shape of a cell using Zernike Moments (see Section 3-2-3) [8, 9, 63]. Since the shape of a malaria parasite is well known, the corresponding features can be identified using Zernike moments and could be used to filter out irrelevant objects. However, instead of first selecting image patches based on intensity and then filter out the objects, these features can be used to implement a filter to directly detect objects with a certain shape, referred to as feature detection.

3-2 Feature Detection

The theory of feature detection is related to how humans recognise objects; feature detection mirrors the brain's early visual processing stage that involves identifying fundamental features like shape, colors and location to identify and classify objects. In other words, this stage establishes *what* is *where* [19]. In case of malaria detection, this means we can use feature detection to detect the parasite together with its location (*where*), and what it looks like (*what*).

In Figure 3-2 a common feature is illustrated, referred to as a corner feature. Features detectors are however not limited to detecting corner features, but can detect shapes such as circles or ellipses as well, as will be discussed in Section 3-2-1. Hence, to detect an object with a certain shape, such as the blob shape of a parasites' nucleus, a feature detector with a similar semantic interpretation could be used. Although blob detectors have been successfully implemented to detect malaria parasites [43], other objects with a similar shape will be detected as well. To mitigate this, the image patch centred at the feature is used to generate

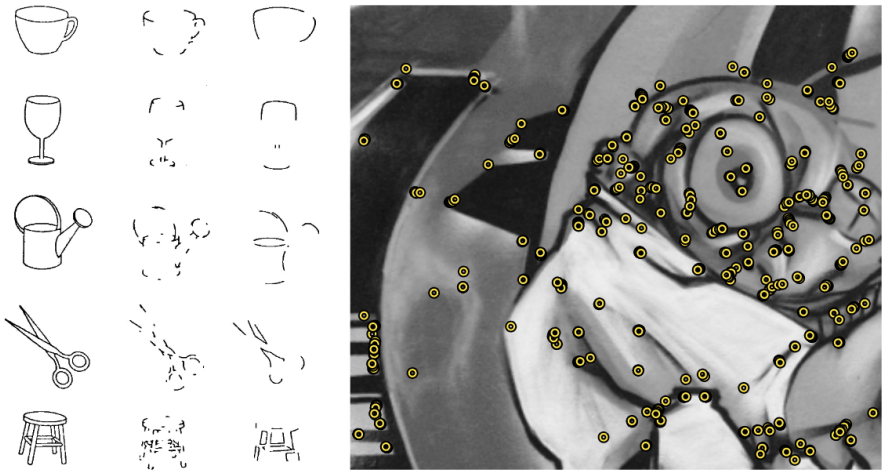


Figure 3-2: Illustration of corner features. Left visualises the effect of removing line segments or corners. On the right, an example is shown of detected corner features in an image. Image adapted from [66]

a description of that part of the image using a feature descriptor. Zernike Moments can be used as feature descriptors, which will be discussed in Section 3-2-3

3-2-1 Feature detectors

Feature detectors can be implemented to find specific features in an image such as shape, often used to find objects in an image that resemble a particular object of interest, such as parasites. Corner, blob and region detectors are common feature detectors in context of feature detection, yet blob detectors are favoured in the medical field [2, 27, 43, 67]. Blob features are typically based on the determinant and/or the trace of the Hessian matrix

$$H = \begin{bmatrix} I_{xx}\left(\mathbf{x}, \sigma_D\right) & I_{xy}\left(\mathbf{x}, \sigma_D\right) \\ I_{xy}\left(\mathbf{x}, \sigma_D\right) & I_{yy}\left(\mathbf{x}, \sigma_D\right) \end{bmatrix},$$

where I_{xx} etc. are the second-order Gaussian derivatives of the image [42, 66]. The trace corresponds to the Laplacian of the image.

$$\nabla^2 L = \text{trace}(H) = \lambda_1 + \lambda_2$$

The eigenvalues are proportional to the curvature in that direction [44]. Therefore, the ratio between the two eigenvalues will indicate if it is a blob feature ($\lambda_1 \approx \lambda_2$), or an edge ($\lambda_1 \gg \lambda_2$). The determinant corresponds to the product of the eigenvalues, and will therefore respond weakly to edges since one of the eigenvalues will be small near edges [14]. Combining the two allows for identifying points that have similar curvatures in both directions [46]. Hence, based on the to-be-detected shape, either a blob detector based on the trace, or the determinant can be used. To derive the size, or scale, of the detected blob feature the scale space is used, which is based on the Laplacian as well.

3-2-2 Scale space

Scale plays an important role in how we perceive the world. Analogously in computer vision scale it is vital for object recognition [42]. The same is true for recognising parasites. However, the scale of a parasite in an image is often unknown. To address this, the scale-space is introduced, in which the proper scale can be selected algorithmically [42]. This allows one to find potential parasites based on an object its scale.

The scale-space representation $L : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ for any two-dimensional signal, $f : \mathbb{R}^2 \to \mathbb{R}$ is defined as the solution to the diffusion equation

$$\frac{\partial L}{\partial t} = \frac{1}{2}\nabla^2 L = \frac{1}{2}\left(\frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2}\right), \quad L(\cdot;0) = f(\cdot) \tag{3-1}$$

where t is the scale parameter. Intuitively this means that as scale increases, peaks in intensity will decrease, i.e., an object seen from a larger distance will get more blurry. Alternatively, the scale-space can be defined as the convolution of various Gaussian kernels $g(x, y; t)$ with $f(x, y)$, referred to as the Gaussian scale-space, defined as:

$$\begin{aligned} L(x, y; t) &= g(x, y; t) * f(x, y) \\ g(x, y; t) &= \frac{1}{2\pi t} e^{-\left(x^2+y^2\right)/2t}. \end{aligned} \tag{3-2}$$

The scale space is normalised such that its magnitude is independent from the scale,

$$\frac{\partial L}{\partial t} = t^{\gamma}\nabla^2 L. \tag{3-3}$$

The scale of an object corresponds to the scale where the magnitude is maximal. Hence, the scale space identifies the scale of an object, which allows matching of objects based on scale, but also allows us to filter out objects that are larger of smaller.

3-2-3 Zernike Decomposition

Feature descriptors are used to give a more detailed description of a feature to distinguish between dissimilar objects that are detected by the same feature detector. Instead of looking for one shape like the feature detector, feature descriptors such as Zernike descriptors incorporate a combination of multiple shapes. Zernike decomposition decomposes an image into a series of orthogonal Zernike polynomials, from which Zernike moments are derived that together describe the shape of an object in said image. Each Zernike moment M_n^m uses a Zernike polynomial V_n^m as a basis function

$$M_n^m = \frac{n+1}{\pi} \sum_x \sum_y \mathcal{I}(x, y) \left[V_n^m(x, y)\right]^*, \quad x^2 + y^2 \leq 1, \tag{3-4}$$

where $\mathcal{I}(x, y)$ denotes the image intensity at (x, y). Zernike polynomials in turn use radial polynomials R_n^m as a basis

$$V_n^m(\rho, \theta) = R_n^m(\rho) \exp(-\mathrm{im}\theta), \tag{3-5}$$

where ρ and θ denote the radius and phase respectively. The following equations are used to transform Cartesian coordinates to $\{\rho, \theta\}$:

$$\begin{aligned} \rho &= \frac{\sqrt{(2x - N + 1)^2 + (2y - N + 1)^2}}{N} \\ \theta &= \tan^{-1}\left(\frac{N - 1 - 2y}{2x - N + 1}\right) \end{aligned} \tag{3-6}$$

The radial polynomials are defined as

$$R_n^m(\rho) = \sum_{s=0}^{(n-m)/2} \frac{(-1)^s (n-s)!}{s!\left(\frac{n+m}{2} - s\right)!\left(\frac{n-m}{2} - s\right)!} \rho^{n-2s} \tag{3-7}$$

Finally, the magnitudes of the Zernike moments can be used as shape descriptors

$$\begin{aligned} I_{n0} &= M_{n0}, \\ I_{nl} &= |M_{nl}|^2 . \end{aligned} \tag{3-8}$$

Hence, the shape of a malaria parasite can be described through Zernike decomposition. The magnitudes can be used to define which basic shapes are dominant, and can therefore be used to detect potential parasite more precisely.

3-3 Conclusion

CNN's are successfully implemented on low-end devices to automatically diagnose malaria when resources are limited, yet its performance deteriorates when applied in-the-field. This drop in performance is reportedly caused by artefacts that are present in in-the-field data, yet due to the CNN black-box architecture this cannot be guaranteed. The hybrid approach proposed in [77–79] is aimed to reduce the complexity, but also opens up the possibility to adapt this algorithm to filter out irrelevant objects to improve precision. However, since most imperfections or errors are different in different datasets, filters tuned on one dataset might not work for other datasets. To circumvent this, one can look for patterns or characteristics that are expected to be constant for all datasets, such as shape. Hence, by implementing a feature detector that matches the expected shape, which can be identified using Zernike decomposition, the overall precision could be improved.

Chapter 4

Dealing with errors and imperfections

Real world data will always have imperfections, and therefore need to be taken into account when automatically diagnosing malaria based on an image of a blood smear. Imperfections can be the result of technical limitations of the imaging setup or inconsistent preparation of blood samples [37], yet the dataset as a whole can be imperfect as well, e.g. the dataset might be too small. The following sections will discuss common error sources, and how to mitigate the negative effect of these imperfections.

4-1 Image quality

4-1-1 Noise

Noise impedes the algorithm's accuracy due to the inherent uncertainty [41]. One of the most effective methods to reduce this effect is median filtering [32, 50]. A median filter assigns the median value of the intensity values of pixels within a window around the current point. The size of the filter defines the window. An advantage is that it does not significantly affect image details, such as the edges of an object. However, if the size of the filter is too large, it can remove small image features [32]. The usual approach to finding the right size is manually changing this value to get the best trade-off between noise reduction and losing details in the image [3, 10]. The noise reduction is typically not quantified but observed manually since there is no ground truth to compare with.

4-1-2 Low image contrast

Low contrast is a common problem when using a brightfield microscope to image biological samples. The result is that only a small portion of the available colour (intensity) scale is used, causing the segmentation and classification process to be hampered [30]. To mitigate this effect, histogram equalisation and contrast stretching techniques are two commonly used

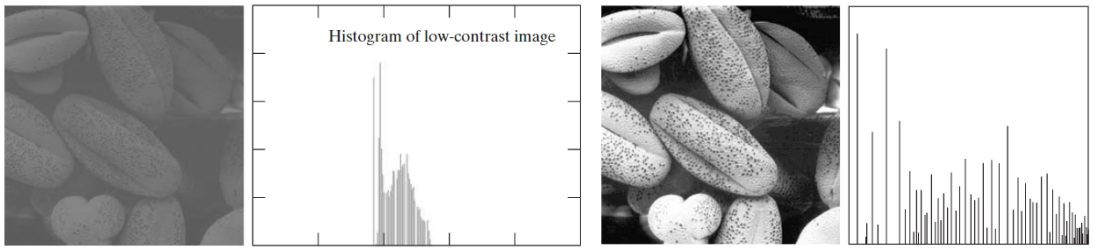


Figure 4-1: Examples of images and their corresponding intensity histograms. (a) a low-contrast image (left), with its corresponding intensity histogram (right), shows that only a small part of the range of possible intensities is used. (b) The same image (left) from as shown in (a) after pre-processing. The histogram (right) shows that the intensities are more uniformly distributed. In this example, histogram equalisation is used to achieve enhanced contrast. Images adapted from [32]

.

methods [24, 50]. To quantify the improvement, measures such as the Peak Signal to Noise Ratio (PSNR) can be used [57, 68]. PSNR is defined as

$$\text{PSNR} = 20 \log_{10} \left(\frac{254}{\sqrt{\text{MSE}}} \right) \tag{4-1}$$

with

$$\text{MSE} = \frac{1}{mn} \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} (\mathcal{I}(x,y) - \hat{\mathcal{I}}(x,y))^2, \tag{4-2}$$

where $\mathcal{I}(x,y)$ and $\hat{\mathcal{I}}(x,y)$ denote the image intensity at (x,y) of the ground truth image, and the deteriorated image respectively. In other words, the further away the intensity of the deteriorated image is to the ground truth the lower your PSNR, i.e, more loss of information. Note that this also means that applying contrast enhancement techniques such as Histogram equalisation (HE) or contrast stretching, as discussed in the next sections, could negatively affect the PSNR. To illustrate this, Figure 4-1b shows a contrast enhanced image, yet differs significantly in intensity from the original (Figure 4-1a), and therefore has a low PSNR. Hence, measuring whether the image quality is improved strongly depends on which metric or quantity is used to quantify it [68].

Histogram equalisation

Low contrast can be mitigated by ensuring that the full range of pixel intensities is used, as visualised in Figure 4-1. This can be achieved through HE. The image is transformed in such a way that the histogram of the pixel intensities in the enhanced image is roughly uniform, i.e., each intensity is represented by approximately the same number of pixels. By creating a uniform histogram, the assumption is that it will give the 'best visual contrast' [30]. Note that this strongly depends on how the contrast is measured. Moreover, this does not directly imply better classification results, since this approach can amplify noise as well [30, 68].

A comprehensive survey of the various variants is presented in [68], which discusses methods to evaluate the performance of these methods as well. [64] and [75] propose that HE can be

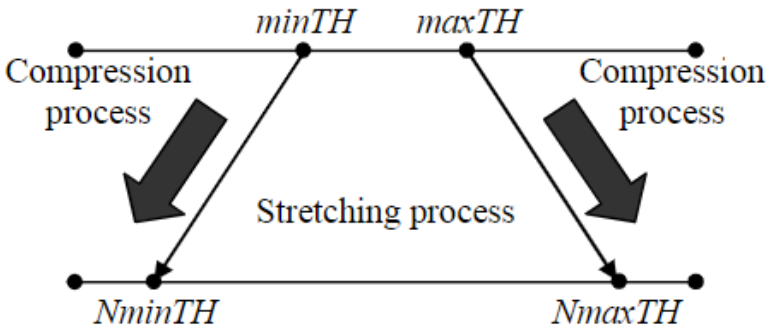


Figure 4-2: Illustration of (partial) contrast stretching. The minimum threshold (minTH) is determined by the minimum non-zero value in an intensity histogram, and vice versa for the maximum threshold (maxTH). These thresholds are then stretched to the desired values. Typically, the minimum is stretched to 0 and the maximum to 255. Alternatively, they can be scaled based on parameters such as standard deviation to achieve partial contrast stretching. Image adapted from [3]

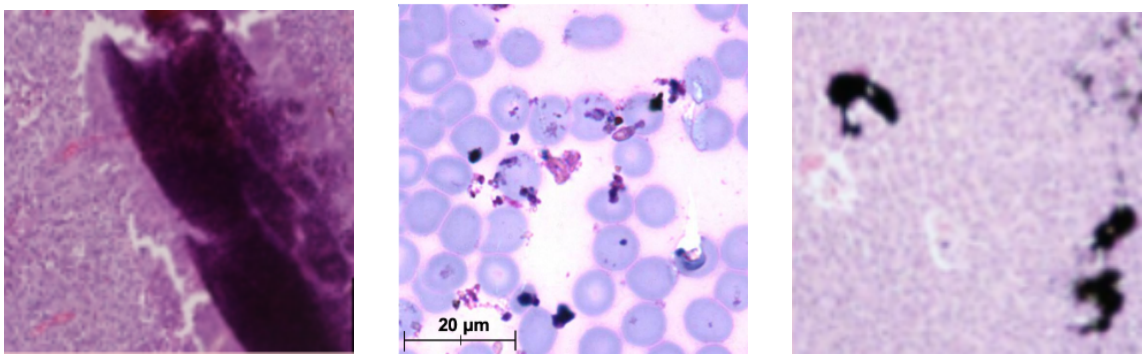
applied to deal with images that are non-uniformly illuminated. The potential improvement of the proposed method in [75] is not evaluated. By using HE for each colour channel, the technique is ensured to work for different detectors and illumination intensities as well [29]. Although this ensures the generalisability to other datasets, it is not clear if it will improve performance for malaria classification for each separate dataset. Moreover, HE algorithms generally suffer from increased computational complexity, brightness reduction, or contrast deficiency [24, 45, 68].

Contrast stretching

To enhance contrast, the lower and upper limits of the image histogram are stretched linearly such that the entire scale is used, as visualised in Figure 4-2. The difference with HE is that whereas HE is designed to get a more uniform histogram (the same number of pixels for each intensity) [32], this method stretches the histogram, and is also less computationally complex. Nevertheless, the threshold are chosen empirically, and can affect the colour balance of the image.

4-1-3 Colour variation

When a sample is not illuminated uniformly, it will result in intensity changes in the image, deteriorating the performance of the consequent image processing tasks [25]. Moreover, due to human inconsistencies in staining a sample, staining variation can occur. This results in discrepancies in colour and intensity. To reduce the effect of intensity changes, grey world colour normalisation is often used [20, 50]. This method assumes that the mean of each colour channel is grey. Hence, channels that have a different mean will be scaled accordingly. Although this method does not take lighting into account [28], making it sensitive to its surroundings, it is an effective method [29]; using different scanners and illuminations, it achieved an average accuracy of 92.3%, whereas the more complex method,HE, only scored



(a) Blotched staining **(b)** Artefacts **(c)** Dirt particles

(d) Dirt particles **(e)** Bacteria

Figure 4-3: Types of artefacts. Adapted from [34, 37, 71]

2% higher. Alternatively, a low-pass filter [11, 25] can be used to reduce this effect. Other approaches include HE (Section 4-1-2).

4-1-4 Artefacts

Artefacts introduced during the acquisition of images can have a wide range of different forms [37]. Depicted in Figure 4-3 are examples of artefacts and contaminants that might be introduced. Although bacteria are not necessarily artefacts, they can lead to confusion due to their similarity to malaria parasites [71]. Besides the fact that these artefacts complicate the diagnosis of malaria, they are not easy to remove or undo either. Small dirt particles can be removed using morphological operations, however, this will remove other small objects that might be of interest as well [37, 50]. Studies have proposed remove images with a significant amount of artefacts from the dataset [34], use a minimum predicted parasite count to account for false positives due to artefacts [78], or use fluorescent microscopy together with brightfield microscopy [33]. Another approach is to train a classifier to recognise these artefacts [38]. However, due to fact that artefacts can basically take any shape or form, it is not trivial to train a classifier for all possible artefacts.

4-2 Data quality

Data quality not only depends on the quality of the images itself, it also depends on the methodology of how these images are acquired. For example, in order to evaluate the per-

formance of the implemented algorithms, the dataset needs to be annotated, i.e., there needs to be a ground truth to compare with. However, there is no standardised way to do this and can therefore differ between datasets. Other factors such as the size of the dataset, and the (im)balance of classes within the datasets play an important role as well. The following sections discuss the effects of these quality factors.

4-2-1 Data size

The performance of classification algorithms depends strongly on the amount of data available [36]. However, it is not trivial to determine how much data is "enough" to properly train your algorithm. One approach is to iteratively evaluate the algorithm for different dataset sizes. The point where the performance stops improving can then be seen as where the dataset is large enough. This approach does, however, assume that the dataset is large enough in the first place in order to see where it stops improving.

From a manual diagnostic perspective the amount of data needed depends on the desired sensitivity for detecting malaria: for lower parasite densities, a higher number of images should be checked to prevent false negatives. Therefore, using the relation between number of images and parasitaemia as discussed in Section 2-3, one can derive if the acquired number of images is sufficient to detect different levels of parasitaemia.

4-2-2 Data annotations

Dataset are annotated manually to establish a ground truth that can be used to evaluate the performance of classification algorithms. Annotations for parasites, or other objects, are typically in the form of a list of coordinates or in counts per image. Whereas the former provides enough information to directly compare the location of the parasite predicted by the model with the ground truth, the latter does not. In other words, if the coordinates are known, the predictions can be defined as True Positive (TP), False Positive (FP), True Negative (TN) or False Negative (FN). These values are required for the performance metrics described in Section 4-3-1. When the coordinates are unknown, evaluation done on image level [78], or by using different metrics such as Pearson's correlation coefficient [33, 34]. Performance metrics based on counts are discussed in Section 4-3-3

Although the main goal of data annotations is to evaluate the performance, it is also essential for analysing the distribution of classes or categories within the data. For example, the data might contain more negative patients than infected patients (data imbalance) or that the data only contains infected patients that have a low parasitaemia, i.e., patients with severe infections are not represented in this dataset.

4-3 Performance metrics

4-3-1 Binary classification

Performance of malaria diagnosis algorithms are typically evaluated based on binary classification metrics. These metrics use a ground truth to compare the classification results; if

the classifier correctly labels the input, it is either *true positive* or *true negative*. If a positive input is classified as negative, it is called a *false negative*, and vice versa for a *false positive*. In [61], a systematic analysis is done by using machine learning to review which metrics are most reliable for each classification task. By analysing how each metric, as defined in (2-5), depends on each result (true positive, false positive, etc.), one can select the right metric for a specific task. For example, if true negatives are not incorporated in the metric, changing the number of true negatives will not change the metric outcome. In case of parasite detection the dataset is usually imbalanced [41], that is, there will likely be more uninfected cells than infected. Because of this, the inability to detect changes in true negatives is favourable since a higher number of true negatives does not necessarily indicate better performance. Precision and recall are examples of measures that possess this property and are therefore preferred when using imbalanced datasets [12].

4-3-2 Area Under Curve (AUC) and Area Under Precision-Recall Curve (AUPRC)

Metrics such as accuracy are evaluated for a single-threshold to divide the predictions into positive and negative classes, yet it is not trivial to choose the right threshold [56]. To mitigate this, Receiver Operating Characteristic (ROC) and Precision-Recall Curve (PRC) plots can be used. Both use recall on one axis and on the other axis specificity and precision, respectively. From these graphs, a single performance metric can be extracted by calculating the area under each curve, resulting in the AUC, and AUPRC. Although the AUC is more common, it can be misleading when the dataset is strongly imbalanced [56], as is generally the case for malaria. For imbalanced datasets, the AUPRC gives a more informative view since it is not affected by the imbalance [56].

4-3-3 Metrics based on counts

Datasets annotated with only parasite counts for each image are not suitable for binary classification metrics, which complicates performance evaluation. If image coordinates for the parasites are not available, the parasite locations predicted by the trained model cannot be validated. Therefore, metrics are proposed to directly compare the predicted parasite count c_p with the true count $c_\top$. One approach is to use measure the linear correlation between the predictions and the ground truth, combined with linear regression. [33, 34, 76]. The latter is used since a high correlation only indicates a strong linear relationship between the predictions and the ground truth, yet does not measure if they are proportional to each other. In contrast to binary classification metrics, however, this method assumes that the absolute error, $|c_\top - c_p|$, will be equivalent for both small and large parasite counts. In addition, the World Health Organisation (WHO) also prescribes a metric that uses relative error instead of the absolute error (see Section 2-5). Hence, although this approach can be used to compare the performance of algorithms, it does diverge from the approaches used in binary classification.

4-4 Conclusion

Methods to improve image quality can be effective for specific datasets, but since many datasets are private [81], these improvements are often not reproducible. Even with public datasets, methods are typically optimised for that specific dataset, limiting their applicability in real-world malaria diagnostics where in-the-field data differs from lab data. Besides image quality, dataset size, annotations and class distributions can affect the performance as well. Methods such as data stratification can be used to ensure data quality, yet do not take away the root cause. Therefore, to reliably evaluate the performance choosing the right performance metric is crucial, with AUC and AUPRC being the preferred metrics. Alternatively, the performance can be evaluated using the method prescribed by the WHO: in 50% of the images the count should be within $\pm 25\%$ of the true count. A performance metric similar to this is Mean Absolute Percentage Error (MAPE). Therefore, to ensure generalisability to in-the-field applications, errors and imperfections are usually not removed but rather taken into account in the data processing pipeline and performance metrics.

Chapter 5

Methods

The goal of this chapter is to provide a roadmap on how the research for this thesis was conducted to reduce the false positive rate for malaria diagnostic algorithms that are applied in-the-field. In the first section (Section 5-1) preliminaries are discussed to provide the reader with an overview of the most relevant concepts and terminology used in this chapter. Next, a brief outline of the roadmap is presented in Section 5-2 as a guideline for how the subsequent sections are connected.

5-1 Preliminaries

The following sections will review the characteristics of malaria parasites, summarise the terminology used in this thesis, and provide an overview of the methods used as benchmark.

5-1-1 Malaria parasite

One of the first stages in the malaria parasite life cycle is the ring-stage, owing its name to its distinctive ring shape. In Figure 2-2b, the key elements of the parasite are shown: the cytoplasm and the nucleus. A nucleus is always present in a parasite, varying from a curved shape to a more rounded one as it ages. Apart from the parasite its morphology and the typical position of the nucleus, it is shown in Figure 2-2a that the parasite has an irregular size and shape ($2 \leq \mathrm{d} \leq 3.7\mu m$) [13]. These variations complicate the detection of parasites [23]. Detection is complicated even further due to the fact that blood smear images typically have low contrast, making it difficult to distinguish between background and parasite [23].

To enhance the contrast a dye is applied to a blood smear which stains each part of the parasite with a distinct, more intense colour [23, 58]; this process is called staining. Due to a lack of standardisation and quality control, artefacts can be introduced during the staining process, such as blurring, contamination due to dirt, or smudges due to incorrect application of the dye [35, 37, 49]. These artefacts are often claimed to be the source of deteriorated in-the-field performance [34, 78].

5-1-2 Terminology

The terminology used in this thesis is summarised in Tables 5-1 and 5-2, which define the meaning of several malaria related terms and machine learning terms respectively.

Table 5-1: Malaria related terms

Term	Description
Image	a camera-captured microscopic image of a blood smear.
Interest point/Keypoint	a point in an image with specific characteristics. E.g., the point in an image where the intensity is minimal.
Image patch	a square, typically small area within an image, centred around an interest point.
Parasitaemia	Number of parasites per μl of blood. Also referred to as parasite density.
Parasite	a malaria parasite. Specifically: A ring-stage Plasmodium Falciparum parasite.

Table 5-2: Machine learning related terms

Term	Description
Data(set)	A set of images or image patches.
Training data	A dataset used to train a model.
Validation data	A dataset used to check the performance while training a model to prevent overfitting to the training dataset.
Test data	A dataset that is kept aside during the training process, and is only used for the final performance evaluation.
Lab data	A dataset consisting of images that are acquired in a laboratory setting.
In-the-field data	A dataset consisting of images that are acquired in an in-the-field setting.
Ground truth	A set of annotations belonging to a dataset that can be used as benchmark, e.g., the true locations of parasites in an image, or the correct *label* of an image.
Label	A label assigns an image or image patch to a predefined category, e.g., positive or negative. These categories are also referred to as *classes*.
Class	A predefined category. In this thesis only two classes are used: positive and negative.
Algorithm	A sequence of computational techniques that aims to fit a model based on a set of inputs, and the desired outputs.

Model	The result of an algorithm. A model is used to make predictions on new data, i.e., images or image patches. The predictions correspond to the probability that the input belongs to the positive class.
Classification	The procedure of assigning a label to an image or image patch by using the predictions of a model.
Computer Aided Diagnostics	An umbrella term for machine learning algorithms that are applied in various medical fields to automate diagnosis.
Feature	A point or area in an image that stands out from its surroundings. A prominent characteristic.
Feature Detector	A filter designed to respond strongly to a specific type of feature.

5-1-3 Benchmark overview

In this thesis the algorithms used in the Malaria Screener [77–79] are replicated in order to verify their results, and to allow further analysis of the deteriorated performance when using in-the-field data. The datasets used to produce the results presented in [77, 78] are publicly available [1], and will be used in this thesis to ensure a fair comparison between the benchmark and the proposed approach. Throughout this thesis, the datasets used in [38, 77] will be referred to as laboratory data (or lab data), and the dataset used in [78] will be referred to as in-the-field data.

Based on these datasets, the benchmark implements two algorithms to analyse the digital images of thick blood smears: First, Iterative Global Minimum Screening (IGMS) is used to preselect a number of patches from the image; this is proposed to reduce the processing time [76]. The workings of this algorithm is summarised in Algorithm 1 Second, a Convolutional Neural Network (CNN) with a reduced number of layers is used to provide equivalent performance, but faster processing [76]. Together these two algorithms form the pipeline referred to (in this thesis) as IGMS-CNN.

5-2 Roadmap

It has been demonstrated that Computer-aided diagnostics (CAD) enhances diagnostic performance by eliminating the human element from the process [41, 50]. Human error, stress, and fatigue all impair manual diagnosis performance [41], yet machine learning algorithms are unaffected by these factors. This apparent superiority, however, is not always true: in contrast to automated diagnosis, the use of in-the-field data has minimal impact on manual diagnosis performance [17, 34, 78]. This suggests that there is a pattern associated with malaria parasites that clinicians can identify and that is constant, independent of the data its source. In fact, clinicians are taught to identify malaria parasites based on their color and shape. These characteristics, also known as features, could therefore be used to automatically

[1] https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html

Algorithm 1 Iterative Global Minimum Screening

procedure IGMS($\mathcal{I}(x,y), j, n, r$)
  $n \leftarrow 400$
  $r \leftarrow 22$
  $\mathcal{I}(x,y) \leftarrow \text{image}(x,y)$
  $j \leftarrow 0$
  RemoveBorder($I(x,y)$)
  RemoveWBCs($I(x,y)$)
  while $j \leq n$ **do**
    SelectMinimum($\mathcal{I}(x,y)$)
    $j \leftarrow j + 1$
  end while
end procedure
function SelectMinimum($\mathcal{I}(x,y)$)
  $x_{min}, y_{min} \leftarrow \min(\mathcal{I}(x,y))$
  patch $\leftarrow \mathcal{I}[x_{min} - r : x_{min} + r, y_{min} - r : y_{min} + r]$
  remove patch from $\mathcal{I}(x,y)$
  return $\mathcal{I}(x,y)$
end function

detect parasites. However, as Figure 2-2 illustrates, these features can differ significantly between samples [71]. Note that although the ring shape of the parasite varies over its lifespan, the *nucleus* is always present [13].

Besides variations between samples, identification of malaria parasite features is complicated even further by the fact that, in contrast to human vision, features are not always invariant to image transformations such as scaling, rotation, and translation [62]. Previous studies have attempted to find features corresponding to malaria, [20–22], yet they did not focus on *why* these features would be relevant for detecting parasites and did not investigate if the features were invariant to image transformations. In other medical fields, however, studies have been able to relate certain features to the shape of a cell using Zernike Moments Although Zernike Moments are by design only invariant to rotation, they can be adapted to attain scale and translational invariance as well [47, 65] Therefore it is proposed to use Zernike moments to identify the dominant shapes in infected image patches and use this information to preselect image patches instead of solely relying on image intensity (which is the current implementation, IGMS). To find image patches that contain these specific shapes a Difference-of-Gaussian (DoG) feature detector is implemented, as will be discussed in Section 5-5.

After implementing the DoG detector, it is used to extract image patches to create a dataset that is used to train and test the benchmark classification algorithm. However, before it is used for trained and testing, the dataset is resampled such that the number of positive and negative patches are equal. A fraction $\eta = \frac{n_{pos}}{n_{neg}}$ of negative patches is sampled from each individual image to ensure that all images are represented in the final dataset. This is referred to as stratification, and will be discussed in Section 5-6-1. Using the created dataset, a new model will be generated from the benchmark classification algorithm, which will be referred to as DoG-CNN. k-fold cross-validation is used during the training process to estimate the accuracy of the model is predictions. This method of cross-validation is discussed in Section

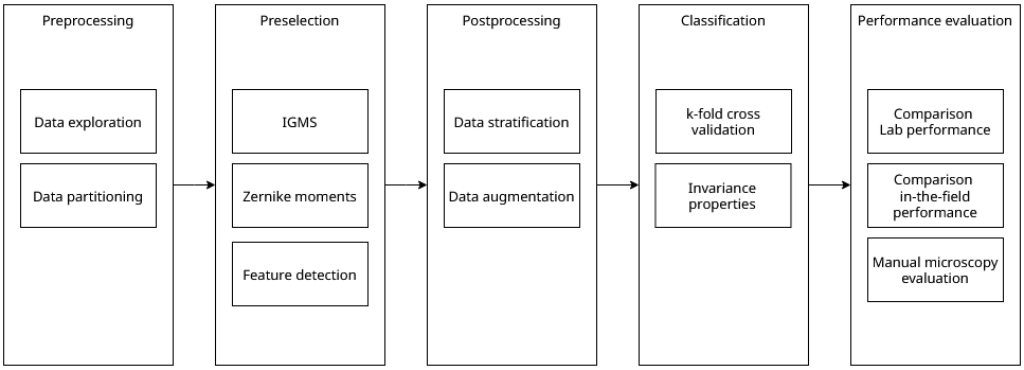


Figure 5-1: Overview pipeline

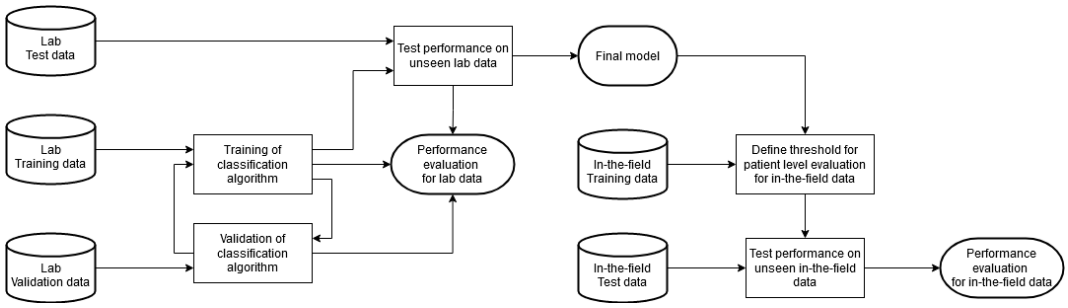


Figure 5-2: Flowchart of evaluation process for both lab and in-the-field data. The training and validation lab data are used to train the classification algorithm, which is then evaluated on unseen data (test data). Next, the in-the-field performance is evaluated. Using the trained algorithm, predictions are made for each sample in the training set. Then, the predictions are averaged for each patient. If the average is above a certain threshold the patient will be classified as infected.

5-7-2.

The last step in the process is performance evaluation. This evaluation is divided into two steps: preselection evaluation; comparing the precision and recall of the DoG detector with the benchmark algorithm IGMS, and classification evaluation; comparing the diagnostic performance of the DoG-CNN with the benchmark model (IGMS-CNN) for both lab and in-the-field data. Classification performance is reported for all metrics shown in Table **??**, from which the Area Under Curve (AUC) and Area Under Precision-Recall Curve (AUPRC) are used to compare the two models. These two metrics are used since they do not rely on one single threshold, and provide insights on how data imbalance affects the performance.

Each of the preceding steps is visualised in Figure 5-1, and will be discussed in more detail in the following sections.

5-3 Data Partitioning

From each dataset only the thick blood smear images that are either negative or contain malaria species *P. Falciparum* are used, since in-the-field performances for thin smears and

other species are not reported [78], and therefore cannot be used for comparison. The lab and in-the-field data will be split into training, validation and test datasets on patient-level [77,78]; that is, each dataset is assigned a predefined number of positive and negative patients (see Table 5-3). In Figure 5-2 a flowchart is shown that depicts how each separate dataset will be used to compare the performance of the benchmark with the algorithm proposed in this thesis.

5-4 Data Exploration

To analyse if the available data is 'enough', the parasite densities are analysed for each patient from which a minimum number of images can be derived, as will be discussed in the next section. Moreover, the calculated parasite densities are categorised into levels of severity, summarised in Table 5-4, such that performance can be evaluated for different densities.

5-4-1 Parasite Density

To minimise the probability of false negatives, the World Health Organisation (WHO) prescribes that, assuming there are 1000 images, at least 100 of them should be analysed before labelling it as negative. This infers that the minimum parasite density that can be detected using this method is 4 parasites/μl, as shown in 2-3. However, to calculate the parasitaemia (2-1), less images need to be analysed [71]: it is prescribed that the technician continues reading slides until 200 White Blood Cells (WBC's) are counted ($\pm$20 images). If at this point the count of parasites, c_p, is not sufficient ($c_p < 100$), counting should be continued until 500 WBC's are counted. Hence, for a reliable calculation of the parasitaemia, the number of WBC's and number of parasites given by the ground truth should conform to these conditions.

To quantify the reliability of the calculated parasitaemia (2-1), the FNR is calculated for each patient. The population size N is assumed to be $N = \frac{500}{20}$, i.e., the approximate number of images needed to count 500 WBC's. n and C_p correspond to the patients count of WBC's and parasites respectively. Using the FNR derived in (2-3) as threshold to ensure the same accuracy, the calculated parasitaemia is labelled as reliable if the $FNR > 34.8\%$ for a given

Table 5-3: Summary of lab and in-the-field datasets in terms of annotations and splits.

Dataset	Annotations		Split	Patients	
	No.	Type		Positive	Negative
Lab	84 509	Bounding boxes	Train	90	30
			Validation	30	10
			Test	30	10
In-the-field	29 034	Counts	Train	21	45
			Test	40	45

n and C_p:

$$\text{FNR} = \left(\frac{500/20 - n}{500/20}\right)^{C_p} > 34.8\% \tag{5-1}$$

For reference, Table 5-5 shows the FNR's for a select range of parasitaemia and numbers of images per patient.

5-5 Adapted preselection

Incorporation of prior knowledge (expert knowledge), instead of solely relying on learned features, is reported to improve generalisability of the algorithm [15, 16]. This infers that incorporating characteristics, such as the shape of a parasite, can improve the algorithm its performance. Although the appearance of malaria parasites is well-documented [71, 73], it is unclear what features are linked to which characteristic of a malaria parasite. Although it is expected that blob features can be used to detect parasites their nucleus due to their similar shape, it is not clear from the literature if this is indeed the case. Since it is shown to be possible to link cell shapes to Zernike moments [8, 9, 63], and the fact that these moments are well suited to be used as features for object recognition due to their well-defined mathematical properties [66], it is proposed to analyse the parasite shapes using Zernike decomposition. The goal of this analysis is to justify the choice to use a blob detector by providing evidence that these detectors can indeed detect parasites. Moreover, after designing the blob detector, Zernike decomposition is used to verify if the detector is more selective towards parasites than IGMS.

5-5-1 Zernike Decomposition

As discussed in Section 3-2-3, the Zernike moments for digital images are defined as:

$$M_n^m = \frac{n+1}{\pi} \sum_x \sum_y \mathcal{I}(x,y) \left[V_n^m(x,y)\right]^*, \quad x^2 + y^2 \leq 1. \tag{5-2}$$

Table 5-4: Degrees of severity for malaria infection in terms of parasitaemia [6, 70]

Severity	#Parasites	Parasitaemia
Low	1–10	4–40
Mild	11–100	44–400
Moderate	101–1000	404–4000
High	1001–25 000	4004–100 000
Hyper	$> 25\,000$	$> 100\,000$

#Parasites corresponds to the number of parasites in 100 images. The range of #Parasites corresponds to a range of integer values. Using the number of parasites, the parasitaemia is approximated using (2-1) where $n_{WBC} = 20$ [71]

To ensure that Zernike moments are not affected by translation, they are centred around the centre of mass $\{\bar{x}, \bar{y}\}$ of an image patch $\mathcal{I}(x, y)$ with size $N \times M$ [47,65]. The centre of mass is defined as

$$\{\bar{x}, \bar{y}\} = \left\{ \frac{\mu_0^1}{\mu_0^0}, \frac{\mu_1^0}{\mu_0^0} \right\}, \tag{5-3}$$

where

$$\mu_j^i = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} x^i y^j \mathcal{I}(x, y). \tag{5-4}$$

However, instead of the image patch centroid, the parasite its centre of mass should be used in order to get the parasite its shape. Hence, before calculating the centre of mass, Otsu thresholding is used to binarize the image patch. By setting the background to zero, only the intensities of the parasite are taken into account, thereby ensuring that the correct centre of mass is selected. An example is shown in Figure 5-3.

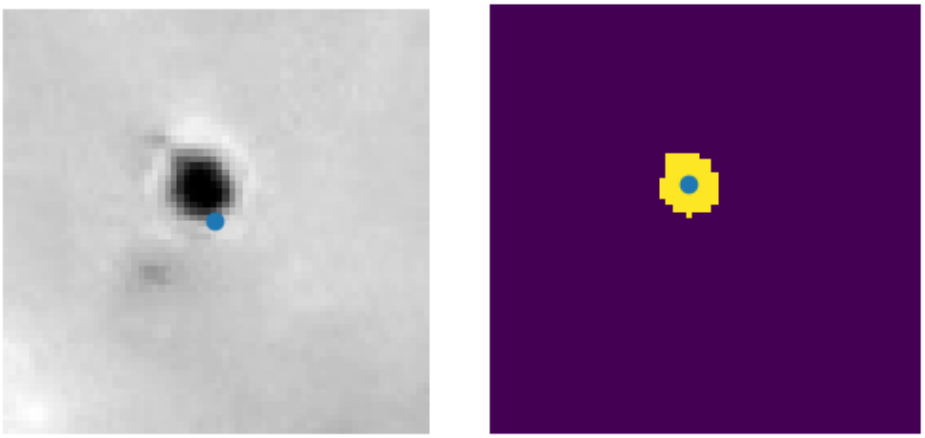
If the radius of the object is known, scale invariance can be attained for Zernike Moments by rescaling the object to a unit circle [47]. Hence, for each image patch the size of the parasite needs to be derived. Although the proposed feature detector can automatically detect the scale of the detected object , IGMS does not. To mitigate this, the radius is derived from the binary image using the image processing package for Python, `skimage` [69].

The last step is to calculate the Zernike moments (5-2) for all positive and negative patches to analyse differences in magnitudes between the classes. Positive patches are based on the ground truth annotations, negative patches are extracted by the IGMS method described in Section 5-1-3. It is expected that in image patches containing parasites the magnitude of defocus, coma, and spherical aberration will on average be higher than others due to their resemblance to the parasites shape. In Figure 5-4 the Zernike polynomials are visualised up to the sixth degree.Students t-test is used to verify if for a particular Zernike moment the difference between the distributions for positive and negative patches is statistically significant ($\rho < 0.05$). Zernike Moments for which the difference is significant will then be used to design the feature detector.

Table 5-5: Probability of a false negative when sampling a number of images n from a population of 25 images for a given parasite count.

		Number of images		
c_p	D	10	15	20
1	16	60%	40%	20%
2	32	36%	16%	4%
4	64	13%	2.6%	0.2%

c_p denotes the number of parasites per n images, where D is the corresponding parasitaemia. The given probabilities are used to check if the number of images is sufficient to reliably detect a given parasite density. For example, if there are 2 parasites present in 25 images, there is a probability of 36% that these are not identified when examining 10 images.



(a) Original example image patch **(b)** Binarized image patch

Figure 5-3: Example of a greyscale image patch containing a parasite. a) shows the original image patch. The blue dot represents the centre of mass. b) shows the binarized image patch with corrected centre of mass.

5-5-2 Feature detector

To detect malaria parasites based on prior knowledge of its shape, a blob detector is implemented. Blob detectors can be tuned to detect shapes ranging from ellipsoids to circular, and are able to automatically detect the scale of an object. The detector shape is tuned such that the shape it detects matches the shape found by Zernike decomposition to improve the precision of the preselection algorithm by filtering out irrelevant objects. Other than the shape, three other parameters are used to filter out objects: minimum scale, maximum scale and minimum intensity. The values for these parameters are found using grid search, while optimising for AUPRC. An initial guess for the minimum and maximum scale is derived using the available annotations, as described in the following section.

The scale of an object can be derived by using a scale-space. This section will give a brief overview of the scale-space, which is described in more detail in Section 3-2-2, and the important parameters that are used to *design* the feature detector. Table 5-6 summarises the relevant parameters to configure the scale-space, including the effect of changing the value of that parameter. Figure 5-5 shows an example of a scale-space with four *octaves*, each consisting of four *levels*. Each level corresponds to a certain scale σ_i, which is derived from the base scale of the image σ_0 [44].

$$\sigma_i = k\sigma_{i-1}, \quad i = \{1, 2, 3, ...\}, \tag{5-5}$$

where k denotes the scale ratio. The value of k is chosen such that the scale doubles for each octave. Therefore, if we have n levels for one octave:

$$\begin{aligned} \sigma_{n-1} = 2\sigma_0 = 2\frac{\sigma_1}{k} &= 2\frac{\sigma_{n-1}}{k^{n-1}} \\ \Downarrow & \\ k^{n-1} &= 2\frac{\sigma_{n-1}}{\sigma_{n-1}} \\ k &= 2^{1/n}. \end{aligned} \tag{5-6}$$

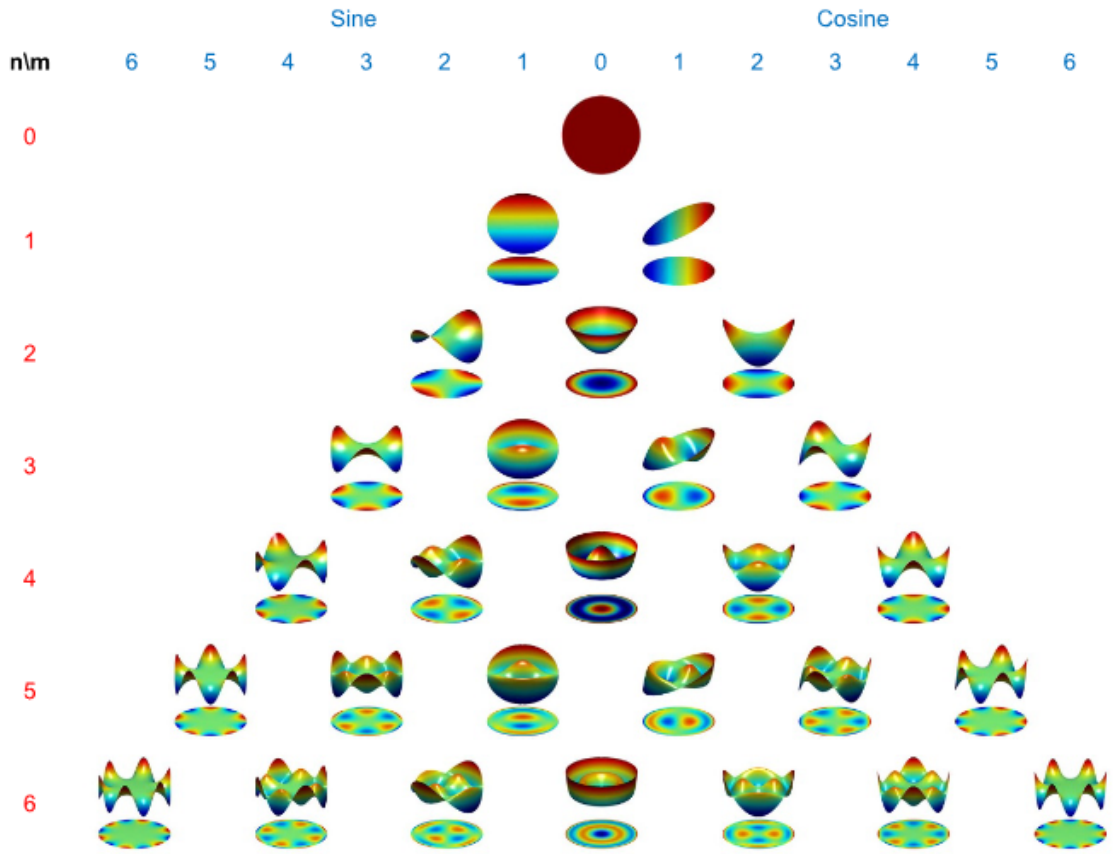


Figure 5-4: Visualisation of Zernike polynomials up to the sixth degree.

Table 5-6: Parameters for configuring a scale-space.

Symbol	Description	Effect
n_o	number of octaves	
n_l	number of levels per octave	More levels will result in a lower amount of blur added per level.
σ	scale	
σ'	relative scale with respect to the previous level	
σ_s	assumed scale at which the image is sampled	Default value is $\sigma_s = 0.5$ [18]
σ_0	base scale of the scale space, $\sigma_0 > \sigma_s$	adds extra smoothing. $\sigma_0 = 1.6$ is recommended [44]
k	scale ratio	

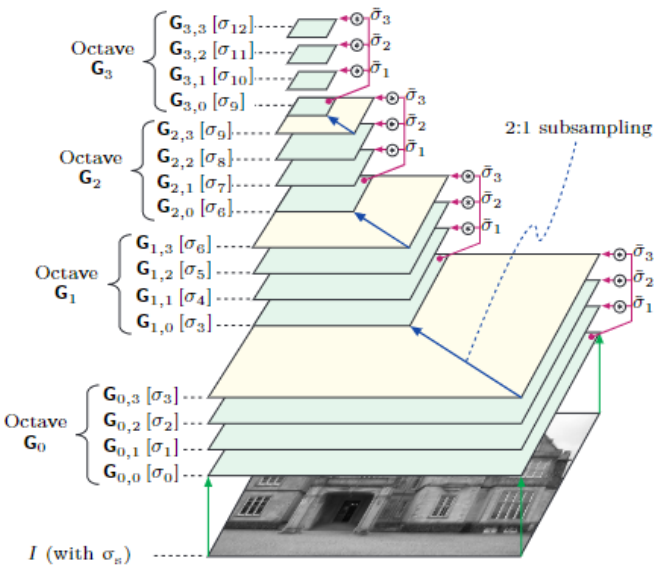


Figure 5-5: Visualisation of a hierarchical scale space. Each octave consists of four levels, where each first level of the next octave is a subsampled copy of the last level of the previous octave. The scale is doubled each octave, which effectively halves the bandwidth. Therefore subsampling will not result in loss of information. Image adapted from [18]

Although a large number of levels per octave infers a better approximation of the scale, it is shown that the accuracy deteriorates for [18, 48]

$$\sigma' = \sqrt{\sigma_n^2 - \sigma_{n-1}^2} < 0.8. \tag{5-7}$$

Therefore, the number of levels n is chosen such that for a given base scale σ_0 the following inequality holds:

$$\begin{aligned} \sqrt{\sigma_1^2 - \sigma_0^2} = \sqrt{k\sigma_0^2 - \sigma_0^2} &> 0.8 \\ \sigma_0^2(2^{1/n} - 1) &> 0.8^2 \\ 2^{1/n} &> \frac{0.8^2}{\sigma_0^2} + 1. \end{aligned} \tag{5-8}$$

To improve the precision of the feature detector the base- and maximum scale are fine-tuned to filter out objects that are smaller or larger than parasites. Although the radius of the parasites is given in the ground truth, it corresponds to the entire parasite, whereas the scale detected by the feature detector might correspond to only a part of the parasite (such as the nucleus) depending on the type of feature detector. Therefore, the ground truth radius is used as initial guess for the base- and maximum scale, and are then fine-tuned to maximise AUPRC

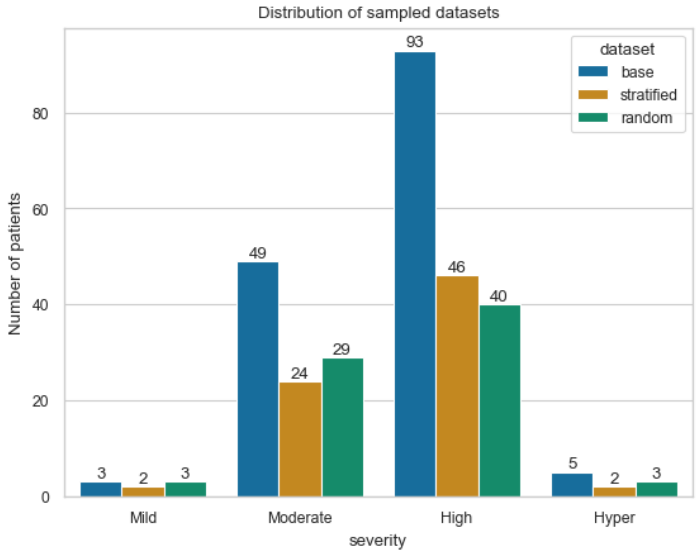


Figure 5-6: Visualisation of sampling a dataset using stratification. The population consists of 150 patients with varying infection severity. In this example, 50% of the dataset is sampled. In contrast to random sampling, stratified sampling ensures that the same fraction (50%) is sampled from each category (severity). This ensures that the sampled dataset is representative to the original dataset.

5-6 Classification

The classification algorithm used in this thesis is a replication of the CNN proposed in [76]. Training of the algorithm is done in a similar way as proposed in the original study: First, image patches are extracted from the images in the training set and are labelled as positive or negative using the provided ground truth. Second, the training set is balanced by using the same number of positive and negative image patches. Finally, the classifier is trained using k-fold cross-validation (as will be discussed in Section 5-7-2). However, instead of taking n_p (number of positive patches) random samples from all negative samples, I propose to use a sampling approach that is shown reduce variance and bias of k-fold cross validation estimate [54]: stratification.

5-6-1 Data Stratification

To ensure that the training and test data are both representative, i.e., that they have a similar distribution of classes as the entire dataset, stratified sampling is applied. Stratification, as exemplified in Figure 5-6, ensures that each 'group' in the data, i.e., patients and images, is represented proportionally. Note that although the example shows stratified sampling based on positive or negative images, stratification can be applied for other groupings as well. This approach is implemented for two separate sampling processes: sampling of patients to form training, validation and test sets, and sampling of patches during k-fold cross validation. The former is achieved by grouping patients by severity, which is defined in Table 5-4. Then, the patients are sampled such that each degree of parasitaemia is represented in each Random Forest [67] (interpretability)

5-7 Evaluation

The evaluation of the DoG-CNN pipeline can be split up into two parts: Lab performance, and in-the-field performance. To evaluate the lab performance, the training set will first be used to fine-tune the parameters of the preselection and classification algorithms. After fine-tuning, the generalisation performance of the classification algorithm is evaluated using the test set. Finally, the predictions are aggregated to derive the performance on patient level. These topics will be addressed in more detail in Section 5-7-1 and 5-7-2. In-the-field performance is evaluated using the aforementioned fine-tuned preselection and classification algorithms. To evaluate the patient level performance, first a threshold is derived that is used to classify a patient as infected or uninfected. This is done in two different ways: the first approach is implemented for comparison with the benchmark, the second approach is implemented to provide a performance evaluation method that is independent from parasitaemia. Section 5-7-3 will discuss these approaches in detail.

5-7-1 Preselection performance

The preselection algorithms their precision and recall are evaluated on image level, and for each severity (see Table 5-4). In contrast to the classifier algorithm, the preselection algorithm does not produce predictions, but rather possible locations of parasites; so called interest points. These interest points are compared with the ground truth annotations $\{x_\top, y_\top, r_\top\}$ to see if they are *close enough* to label it as positive. Specifically, the interest point $\{\hat{x}, \hat{y}\}$ is counted as True Positive (TP) if:

$$\sqrt{x_\top^2 + y_\top^2} - \sqrt{\hat{x}^2 + \hat{y}^2} \leq r_\top, \tag{5-9}$$

where $r_\top$ is the radius of the parasite according to the ground truth. If (5-9) does not hold, it is counted as False Positive (FP). Ground truth annotations that have no interest points close to them are counted as False Negative (FN).

For each image, precision and recall are calculated using the aforementioned quantities. From these results the 95% Confidence Interval (CI) is calculated as final metric. This interval is calculated by taking the 2.5 percentile as lower limit, and the 97.5 as upper limit. To analyse potential correlations between the performance metrics and parasitaemia, the performance will be reported for each parasitaemia level, as defined in Section 5-4-1.

5-7-2 Classification performance

A CNN is used to automatically diagnose if a blood smear image is infected or not, i.e., the CNN is used to *classify* the images. Classification refers dividing the dataset into *classes* through the assignment of categorical labels to samples from the dataset. In this case, there are two classes: positive and negative. Table 5-7 summarises some common labels analogous aforementioned classes. A label is assigned to a sample based on the prediction score of the CNN: if the prediction score exceeds a certain threshold γ_b it is considered positive, and negative otherwise. Unless specified otherwise, this threshold is set to $\gamma_b = 0.5$. Before discussing which performance metrics are used, the next paragraph elaborates on *how* the model is evaluated.

Positive	Negative	Used in
Infected	Uninfected	Diagnosis
1	0	Binary classification

Table 5-7: Analogous terminology for positive and negative classes.

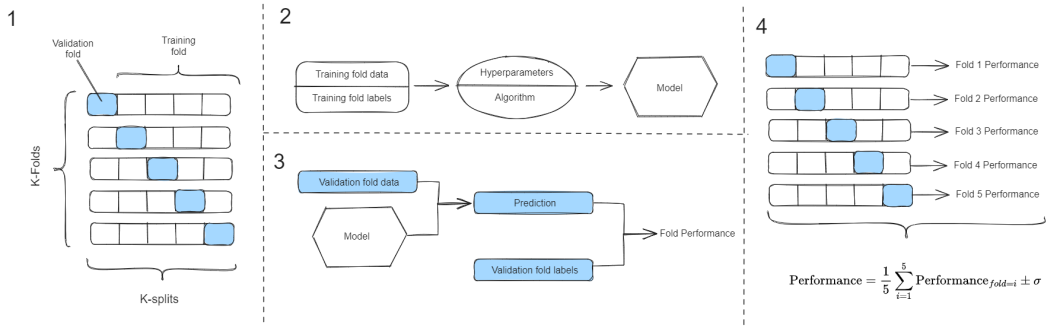


Figure 5-7: Visualisation of k-fold cross-validation procedure. The first step is splitting the dataset into k-splits, in this case $k = 5$. From these splits, five distinct folds are created. Note that the training data has some overlap between the separate folds, whereas the validation data does not. The second step is to train a model for each fold, which are then used in the third step to evaluate the performance of each fold. The last step is to calculate the final performance by taking the average and standard deviation (σ) over all fold performances. Visualisation is based on the approach described in [54]

The classification model is evaluated using k-fold cross-validation, which reduces the bias of the estimated performance by using a larger portion of the dataset as training data. Instead of using a separate training and test sets, the dataset is divided into k-splits. The training process is repeated k times, where each *fold* uses $k-1$ splits for training and one for validation. Figure 5-7 visualises how each fold is composed using different sets of splits. One of the main advantages are that this method reduces overfitting, and does not waste as much data compared to splitting the dataset into three parts, since each split can be used for either training and validation [54]. To ensure a fair comparison between the benchmark and the proposed method, the same number of folds ($k = 5$) is used [77]. The final performance is derived by taking the mean and standard deviation of the performances for each fold.

Using the results from k-fold cross validation, the performance metrics defined in Table **??** are calculated to compare between the benchmark and the proposed pipeline. Although all metrics are calculated to provide a comprehensive comparison, only the AUC and AUPRC are used to analyse if the proposed method improves upon the benchmark. These metrics are selected since they both take into account how the model performs for different thresholds instead of one. The choice to use both metrics is based on the fact that the AUC is used for evaluating the benchmark [77, 78], whereas the AUPRC is shown to be more informative than the AUC in case of imbalanced datasets [56].

The last step is to aggregate the results on patch level to patient level. First, the performance is evaluated for each image. Second, the patient level performance is derived by calculating the 95% CI from the image results for each patient.

5-7-3 In-the-field performance

In-the-field performance is evaluated, if possible, in the same way as the lab performance. The evaluation deviates from the previous evaluations in that it does not evaluate on patch level, since the in-the-field data only contain parasite counts for each image instead of locations. Moreover, the threshold γ_f used to label a prediction as positive or negative is derived in two different ways, as discussed in the next paragraph.

To allow comparison with the benchmark the first approach is based on [78]: First the average of all predictions for each patient is calculated, from which a threshold is derived that *'maximises'* the performance [78]. Since it is not reported what specific metric or loss function is maximised, the AUC and AUPRC will be used in the comparison. This eliminates the need of setting a fixed threshold that is tuned for a specific metric. The prediction score of the j^{th} image, $\bar{p}_j$, is the average of all m predictions $p_{i,j}$ belonging to image j:

$$\bar{p}_j = \frac{1}{m}\sum_{i=1}^{m} p_{i,j}. \tag{5-10}$$

Similarly, the prediction score of the k^{th} patient, $\bar{P}_k$, is the average of all n image predictions scores $\bar{p}_{i,k}$ belonging to patient k:

$$\bar{P}_k = \frac{1}{n}\sum_{i=1}^{n} \bar{p}_{i,k}. \tag{5-11}$$

Note that since the number of predictions is equal to the number of image patches extracted from the images, and the IGMS algorithm is designed to extract a fixed number of patches, the image prediction score depends on the number of parasites present in the image. Specifically, if the number of patches correctly identified as positive is high, this infers that the average prediction score is high as well. Due to this dependency, it is expected that when fine-tuning this threshold to a dataset with a relatively low parasite density, it will increase the probability of false positives when testing on datasets with higher parasite densities and vice-versa.

The second approach defines γ_f as the threshold for which the Mean Absolute Percentage Error (MAPE) between predicted number of parasites $\hat{C}(\gamma_f)$ in an image and actual number of parasites $C_\top$ is minimal. This eliminates the dependency on parasite densities and makes it consistent with the performance evaluation approach prescribed by the WHO (See Section 2-5).

$$\begin{aligned} \min_{\gamma_f} \; & 100\frac{1}{k}\sum_{j=1}^{k}\left|\frac{C_{\top,j} - \hat{C}_j(\gamma_f)}{C_{\top,j}}\right| \\ \hat{C}_j(\gamma_f) &= \#(\{p_{i,j} \in \mathbf{p}_j \mid p_{i,j} > \gamma_f\})). \end{aligned} \tag{5-12}$$

$\mathbf{p}_j$ denotes the set of predictions belonging to image j, $\#(\cdot)$ denotes the number of elements (cardinality).

5-8 Summary

Together, the methods described in the previous sections form the proposed malaria diagnostic pipeline of this thesis. This pipeline is designed to address the following issues to reduce the

false positive rate for in-the-field applications: First, the preselection algorithm is adapted to take the shape of the parasite into account such that irrelevant objects can be filtered out. Second, the data (image patches) extracted by the preselection algorithm are partitioned into training and validation sets while taking parasitaemia into account to ensure that each dataset is representative for the original dataset. Third, the performance evaluation is adapted to take into account data imbalance by using the AUPRC. Lastly, to allow evaluation based on parasite counts rather than locations, the MAPE is used, which also provides a metric that is equivalent to the one used to evaluate the performance of microscopists. Hence, this pipeline adapts the algorithm to reduce the false positive rate, adapts the methodology of performance evaluation to provide a more informative view when data is imbalanced, and implements MAPE to reliably evaluate in-the-field performance.

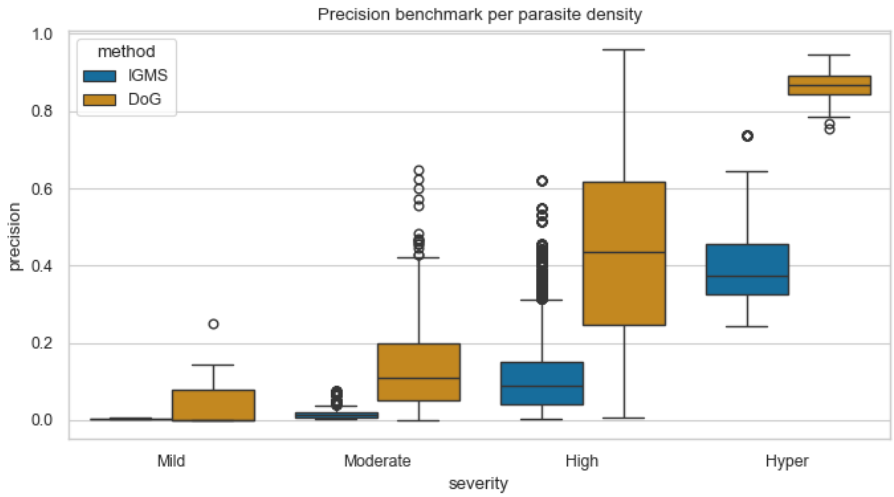
Chapter 6

Results

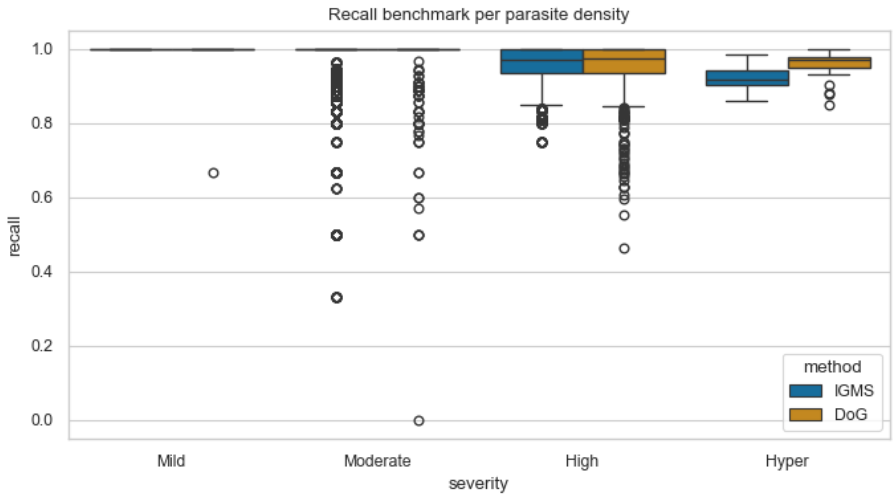
In this chapter the preselection performance of the proposed Difference-of-Gaussian (DoG) blob detector are benchmarked against the Iterative Global Minimum Screening (IGMS) algorithm, and evaluated how it affects the performance of the DoG-CNN and IGMS-CNN models for both laboratory and in-the-field data. By implementing the DoG detector tailored to detect shapes similar to malaria parasites, it is expected that the precision of the preselection algorithm will be improved compared to IGMS. As a result, it is expected that the DoG-CNN performance will improve as well. To analyse other possible factors that negatively affect the performance, the effect of data stratification on classification performance of the DoG-CNN pipeline is evaluated. By implementing data stratification is not necessarily expected that the mean performance will improve, yet the variance should be lower.

6-1 Improved preselection precision

The DoG feature detection algorithm is fine-tuned to only select objects that have a similar size and shape as parasites to filter out irrelevant objects while maintaining recall, which in some cases halves the amount of false positives when compared to IGMS. As shown in Figure 6-1, the recalls of DoG and IGMS overlap for mild to high parasitaemia, however, the mean value of DoG for hyperparasitaemia is 4% higher. As for precision, the mean values of DoG are higher (4%, 13%, 33%, 46%) for all parasitaemia, yet the DoG has a larger interquartile range (15%, 26%) for moderate and high severities.



(a)



(b)

Figure 6-1: Performance comparison of IGMS algorithm and the feature selection algorithm DoG, evaluated on image-level for lab data.

Hence, the DoG feature detector is able to maintain an equivalent recall as the benchmark, while improving the mean precision for all parasitaemia. Also, the mean recall and precision for hyperparasitaemia, 96.14% and 86.6% respectively, are noteworthy considering the uncomplicated nature of the feature detection algorithm. The IGMS algorithm does have a smaller spread for all but hyperparasitaemia. However, this could be explained by the fact that IGMS extracts a fixed number ($n = 400$) of image patches for each image. If all parasites are found, recall $=100\% \Rightarrow \text{TP} = c_p$, $\text{FN} = 0$, and we have 400 patches predicted as positive

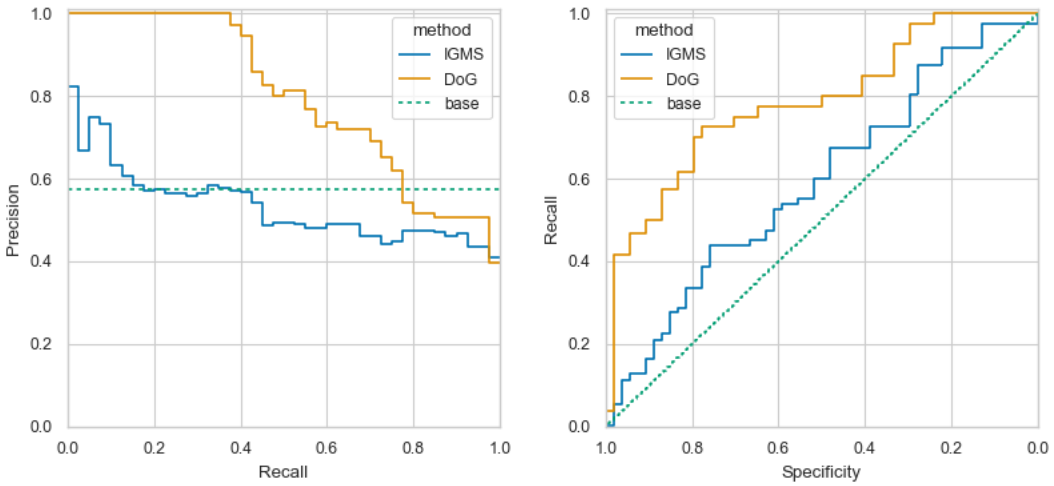
(PP), it follows that:

$$\begin{aligned} \text{PP} = 400 = \text{FP} + \text{TP} \rightarrow \text{FP} = 400 - c_p \\ \Rightarrow \text{precision} = \frac{c_p}{c_p + (400 - c_p)} = \frac{c_p}{400} \end{aligned} \tag{6-1}$$

Hence, the precision is upper bounded, which limits its spread.

6-2 Comparison of in-the-field classification performance

Using the feature detector DoG proposed in this thesis, the classification performance is improved in terms of Area Under Precision-Recall Curve (AUPRC) (26.4%), and Area Under Curve (AUC) (19.4%) when compared to IGMS. Although both classifiers have a higher recall than the baseline for all specificities as shown in Figure 6-2b, the precision is worse than the baseline (<57.5%) for a recall higher than 17.5% and 77.5% for IGMS and DoG respectively, as shown in Figure 6-2a.



(a) Precision-Recall Curve **(b)** ROC Curve

Figure 6-2: Performance evaluation of classifiers for in-the-field data. The classifiers are trained using image patches extracted by IGMS and DoG from lab image data. For the Precision-Recall plot, the baseline is derived from the class imbalance: $n_{c=1}/(n_{c=0} + n_{c=1})$, where $n_{c=i}$ denotes the number of samples for class i. The classifier trained and tested on DoG data has a 26.4% and 19.4% larger area under the Precision-Recall Curve (PRC) and the Receiver Operating Characteristic (ROC) curve respectively.

6-3 Effect of data stratification on classification performance

It is found that by ensuring that the both training and validation datasets have an equivalent distribution of positive and negative classes, and infection severity per patient, the standard deviation of the validation loss is less than half ($\sigma = 30.8 \times 10^{-3}$, compared to $\sigma = 68.1 \times 10^{-3}$) than without stratification. Figure 6-3 visualises the loss during training of the models based on the DoG-CNN pipeline, with on the left the result for without stratification, and on the

right with stratification. Whereas the training loss is equivalent for both approaches and keeps decreasing, the validation loss varies for each epoch and increases in the last epochs. This could indicate that the models are overfitting. Hence, stratification results in a lower deviation in validation loss, which indicates that the model generalises better to new, unseen data.

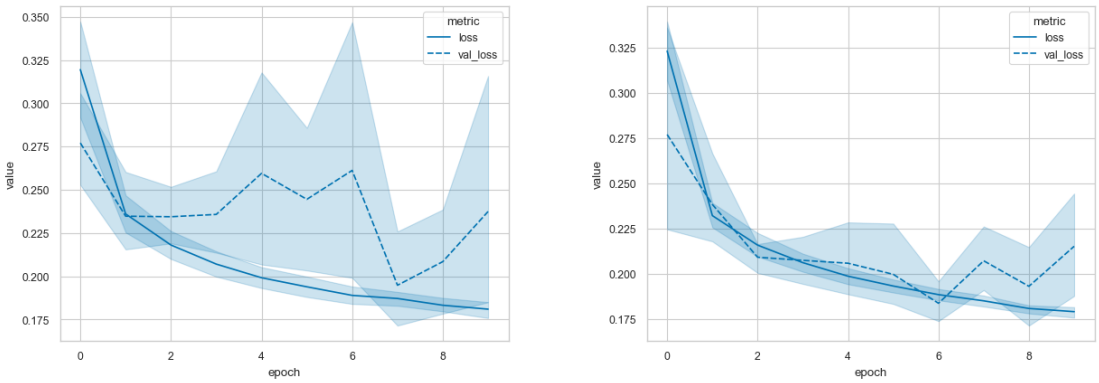


Figure 6-3: Training and validation loss over 10 epochs without stratification (a), and with stratification (b). Both graphs plot the results of 5 models resulting from k-fold cross-validation. The average training loss is equivalent (0.18), and is relatively stable across the different models in terms of standard deviation ($\sigma = 5.7 \times 10^{-3}$ and $\sigma = 4.6 \times 10^{-3}$ for without and with stratification respectively). The deviation in validation loss is highest when no stratification is applied ($\sigma = 68.1 \times 10^{-3}$, compared to $\sigma = 30.8 \times 10^{-3}$).

6-4 Identification of parasite shapes using Zernike Decomposition

By decomposing the images it is shown that on average, *defocus* and *spherical aberration* (corresponding to noll-indices $j = \{4, 11, 22, 37\}$) differ the most in terms of magnitude. These Zernike moments all correspond to the shape of a blob detector, and therefore supports the choice of using a blob detector to detect malaria parasites. Table 6-1 specifies the difference in magnitude with negative patches for these moments.

Table 6-1: Differences in magnitude for positive and negative patches

Label	Defocus ($j = 4$)	Spherical aberrations		
		Primary ($j = 11$)	Secondary ($j = 22$)	Tertiary ($j = 37$)
Negative	0.096×10^{-3}	0.042×10^{-3}	0.046×10^{-3}	0.025×10^{-3}
Positive	0.184×10^{-3}	0.170×10^{-3}	0.187×10^{-3}	0.132×10^{-3}
Difference	0.088×10^{-3}	0.128×10^{-3}	0.141×10^{-3}	0.107×10^{-3}

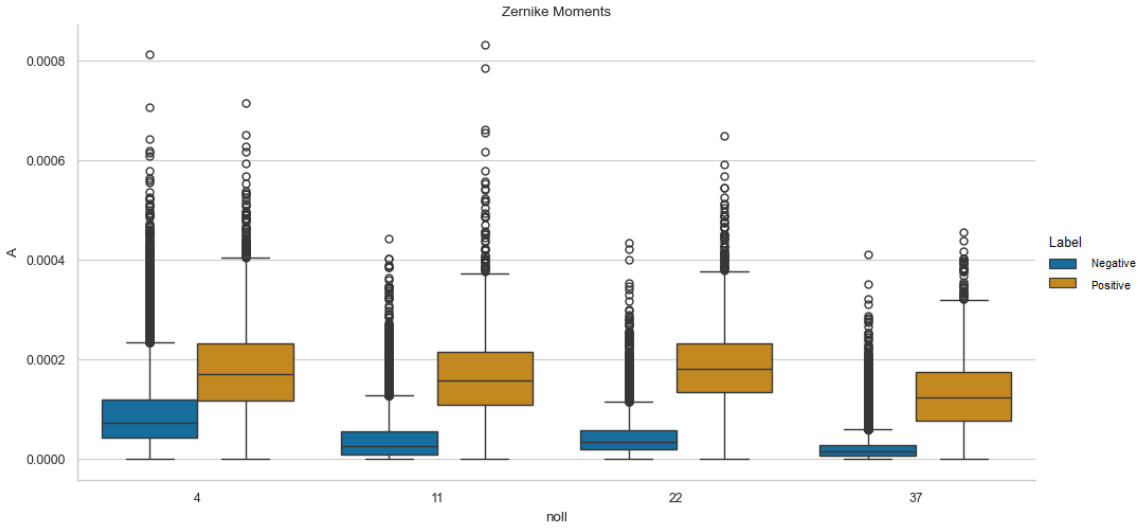


Figure 6-4: Zernike moments for positive and negative image patches extracted from lab data. The noll-indices (j) are plotted on the x-axis, and their corresponding magnitudes on the y-axis. The Zernike moments correspond to defocus and spherical aberrations ($j = \{4, 11, 22, 37\}$).

6-5 Conclusion

The proposed DoG blob detector is shown to be, on average, more precise than the IGMS algorithm in terms of preselecting patches that might contain parasites. This effect is also visible in the performance of the classification algorithm: The DoG-CNN is has a higher mean precision, as shown in the AUPRC plot in Figure 6-2a. From this it can be inferred that, as proposed in this thesis, implementing a feature detector based on the shapes identified through Zernike decomposition can indeed improve the precision, while maintaining recall. Moreover, the AUPRC shows that for a recall higher than 17.5% and 77.5% for IGMS and DoG respectively, the precision of these algorithms is less than if one would let a coin flip decide the diagnosis instead. This effect is not visible in the ROC curve, which infers that the AUPRC gives a more informative view of the performance.

Chapter 7

Conclusion

This thesis aimed to design an automated malaria diagnosis pipeline with improved in-the-field performance by reducing the number of false positives. The pipeline, referred to as DoG-CNN, was designed to diagnose if a patient is infected with malaria by analysing a blood smear image acquired through a microscope, regardless of where the images are acquired. The research questions for this thesis were as follows:

1. How can we identify the sources of false positives generated by state-of-the-art algorithms?
2. How can we reduce the effect of these error sources in order to minimise false positives?

Staining artefacts are reported to be one of the root cause of performance deterioration in malaria detection algorithms. However, since these studies implement a black-box classifier such as Convolutional Neural Networks (CNN's), it is difficult, if not impossible, to deduce if this is indeed the case. The first step was to replicate the benchmark, which showed that the in-the-field performance is indeed lower compared to lab data. To mitigate this, a Difference-of-Gaussian (DoG) blob detector to detect shapes similar to that of malaria parasites, which were identified using Zernike decomposition. While maintaining a similar recall as Iterative Global Minimum Screening (IGMS), the DoG detector improved the precision for all parasite densities by 4.45%, 13.02%, 32.91%, 46.17% for mild, moderate, high, and hyperparasitaemia respectively.

As expected, improving the precision of the preselection algorithm improved the precision of the pipeline (DoG-CNN) as a whole as well. Compared to IGMS-CNN, the mean Area Under Precision-Recall Curve (AUPRC) and Area Under Curve (AUC) are improved by 26.4% and 19.4% respectively. Nevertheless, the average in-the-field performance of DoG-CNN (AUC: 86.8%, AUPRC: 61.7%) remains inferior to the laboratory benchmark performance (AUC: 99.8%, AUPRC: 98.3%). Note that the AUPRC drops by 36.6%, whereas the AUC drops by 13%. Hence, the performance is improved by using DoG, yet the AUPRC exposes a different source of complications: data imbalance.

It was found that not only the image quality affected the results, but also the quality of the data itself as well. In particular, data imbalance in terms of classes (positive/negative) and infection severity (mild, medium, high, hyper) resulted in a higher variance when cross validating models. By applying data stratification to ensure that each severity was represented equivalently in each dataset, the standard deviation was reduced by 55%. Since the available lab data consisted of 150 positive and 50 negative patients, the class imbalance remains. Therefore, AUPRC was used as additional metric since it gives a more informative representation of the performance. Still, this metric was based on the evaluation methodology of the benchmark, which averaged the prediction scores to diagnose a patient. It was found that this metric depends on the average infection severity of patients present in the dataset which varies between datasets, and is therefore not able to generalise to other datasets.

The proposed DoG-CNN pipeline and methodology of performance evaluation have been shown to improve the performance compared to IGMS-CNN, and provides a framework to reliably evaluate in-the-field performance of automated malaria algorithms. The shapes identified by Zernike decomposition can be effectively used to implement a DoG blob detector that is selective towards parasites, improving the precision while maintaining recall. Data stratification, AUPRC provide additional measures to deal with data imbalance. Through these improvements, this thesis paves the way for applying automated diagnostics in resource-limited settings, ultimately aimed to mitigate the burden of malaria in endemic regions.

The application of Zernike Decomposition in this context represents a novel approach to addressing the challenge of shape recognition in malaria parasites, which has not been extensively explored in prior research towards malaria detection. This not only improves the diagnostic algorithm but also provides a framework that can be adapted to use other classification algorithms that use the Zernike descriptors as input, and allows to configure the preselection algorithm to adapt to image transformations such as scaling or rotating.

In alignment with the research questions outlined at the start of this thesis, this thesis has demonstrated how prior knowledge can be incorporated in the preselection algorithm to reduce positives, and provides techniques and metrics to reduce the effect of errors and imperfections in a quantifiable manner. Future research could investigate other datasets that were not present at the time of writing to further evaluate the generalisability of this approach. Moreover, the methodology prescribed by the World Health Organisation (WHO) to evaluate the performance of clinicians could be mirrored by implementing Mean Absolute Percentage Error (MAPE), which has the advantage that it is independent of the average infection severity and is therefore able to generalise better to other datasets.

Bibliography

[1] Convolutional Neural Networks: 1998-2023 Overview | SuperAnnotate.

[2] Adeel Ahmed Abbasi, Lal Hussain, Imtiaz Ahmed Awan, Imran Abbasi, Abdul Majid, Malik Sajjad Ahmed Nadeem, and Quratul-Ain Chaudhary. Detecting prostate cancer using deep learning convolution neural network with transfer learning approach. *Cognitive Neurodynamics*, 14(4):523–533, August 2020.

[3] Aimi Salihah Abdul Nasir, Mohd Mashor, and Zeehaida Mohamed. Colour Image Segmentation Approach for Detection of Malaria Parasites Using Various Colour Models and k-Means Clustering. *WSEAS Transactions on Biology and Biomedicine*, 10:41–55, January 2013.

[4] F. Abdurahman, K.A. Fante, and M. Aliy. Malaria parasite detection in thick blood smear microscopic images using modified YOLOV3 and YOLOV4 models. *BMC Bioinformatics*, 22(1), 2021. 47 citations (Crossref) [2023-12-06].

[5] Aliyu Abubakar, Mohammed Ajuji, and Ibrahim Usman Yahya. DeepFMD: Computational Analysis for Malaria Detection in Blood-Smear Images Using Deep-Learning Features. *Applied System Innovation*, 4(82):82, October 2021. 15 citations (Crossref) [2023-12-06] Publisher: MDPI AG.

[6] Mohamed Al-Salahy, Bushra Shnawa, Gamal Abed, Ahmed Mandour, and Ali Al-Ezzi. Parasitaemia and Its Relation to Hematological Parameters and Liver Function among Patients Malaria in Abs, Hajjah, Northwest Yemen. *Interdisciplinary Perspectives on Infectious Diseases*, 2016:5954394, 2016.

[7] Redha Ali, Russell C. Hardie, Barath Narayanan Narayanan, and Temesguen M. Kebede. IMNets: Deep Learning Using an Incremental Modular Network Synthesis Approach for Medical Imaging Applications. *Applied Sciences*, 12(5500):5500, May 2022. Publisher: MDPI AG.

[8] Elaheh Alizadeh, Samanthe Merrick Lyons, Jordan Marie Castle, and Ashok Prasad. Measuring systematic changes in invasive cancer cell shape using Zernike moments. *Integrative Biology*, 8(11):1183–1193, November 2016.

[9] Elaheh Alizadeh, Wenlong Xu, Jordan Castle, Jacqueline Foss, and Ashok Prasad. TISMorph: A tool to quantify texture, irregularity and spreading of single cells. *PLOS ONE*, 14(6):e0217346, June 2019. Publisher: Public Library of Science.

[10] D. Anggraini, A.S. Nugroho, C. Pratama, I.E. Rozi, V. Pragesjvara, and M. Gunawan. Automated status identification of microscopic images obtained from malaria thin blood smears using Bayes decision: A study case in Plasmodium falciparum. pages 347–352, 2011.

[11] J. E. Arco, J. M. Górriz, J. Ramírez, I. Álvarez, and C. G. Puntonet. Digital image analysis for automatic enumeration of malaria parasites using morphological operations. *Expert Systems with Applications*, 42(6):3041–3047, April 2015. 60 citations (Crossref) [2023-12-06].

[12] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, Tinne Tuytelaars, Jiri Matas, and Krystian Mikolajczyk. \mathbb HH-Patches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2825–2841, November 2020. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[13] Lawrence Howard Bannister, John Mervyn Hopkins, Gabriele Margos, Anton Richard Dluzewski, and Graham Howard Mitchell. Three-Dimensional Ultrastructure of the Ring Stage of Plasmodium falciparum: Evidence for Export Pathways. *Microscopy and Microanalysis*, 10(5):551–562, October 2004. Publisher: Cambridge University Press.

[14] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008.

[15] N. Bento, J. Rebelo, M. Barandas, A.V. Carreiro, A. Campagner, F. Cabitza, and H. Gamboa. Comparing Handcrafted Features and Deep Neural Representations for Domain Generalization in Human Activity Recognition. *Sensors*, 22(19), 2022.

[16] Nuno Bento, Joana Rebelo, André V. Carreiro, François Ravache, and Marília Barandas. Exploring Regularization Methods for Domain Generalization in Accelerometer-Based Human Activity Recognition. *Sensors*, 23(6511):6511, July 2023. Publisher: MDPI AG.

[17] P. Berzosa, A. De Lucio, M. Romay-Barja, Z. Herrador, V. González, L. García, A. Fernández-Martínez, M. Santana-Morales, P. Ncogo, B. Valladares, M. Riloha, and A. Benito. Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea. *Malaria Journal*, 17(1), 2018. 133 citations (Crossref) [2023-12-06].

[18] Wilhelm Burger and Mark J. Burge. Scale-Invariant Feature Transform (SIFT). In Wilhelm Burger and Mark J. Burge, editors, *Digital Image Processing: An Algorithmic Introduction*, Texts in Computer Science, pages 709–763. Springer International Publishing, Cham, 2022.

[19] Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and where: A Bayesian inference theory of attention. *Vision Research*, 50(22):2233–2247, October 2010.

[20] Dev Kumar Das, Madhumala Ghosh, Mallika Pal, Asok K. Maiti, and Chandan Chakraborty. Machine learning approach for automated screening of malaria parasite using light microscopic images. *Micron*, 45:97–106, February 2013. 169 citations (Crossref) [2023-12-06].

[21] D.K. Das, A.K. Maiti, and C. Chakraborty. Automated system for characterization and classification of malaria-infected stages using light microscopic images of thin blood smears. *Journal of Microscopy*, 257(3):238–252, 2015. 39 citations (Crossref) [2023-12-06].

[22] Ishan R. Dave. Image analysis for malaria parasite detection from microscopic images of thick blood smear. volume 2018-January, pages 1303–1307, 2018. 12 citations (Crossref) [2023-12-06].

[23] C.M. de Korne, L. van Lieshout, F.W.B. van Leeuwen, and M. Roestenberg. Imaging as a (pre)clinical tool in parasitology. *Trends in Parasitology*, 39(3):212–226, 2023.

[24] Krishna Gopal Dhal, Arunita Das, Swarnajit Ray, Jorge Gálvez, and Sanjoy Das. Histogram Equalization Variants as Optimization Problems: A Review. *Archives of Computational Methods in Engineering*, 28(3):1471–1496, May 2021.

[25] Gloria Díaz, Fabio A. González, and Eduardo Romero. A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. *Journal of Biomedical Informatics*, 42(2):296–307, April 2009. 155 citations (Crossref) [2023-12-06].

[26] Ben Earnest. Getting Started with Convolutional Neural Networks (CNN) | by Ben Earnest, M. Sc. | AI Mind.

[27] Shahab Ensafi, Shijian Lu, Ashraf A. Kassim, and Chew Lim Tan. Accurate HEp-2 cell classification based on sparse bag of words coding. *Computerized Medical Imaging and Graphics*, 57:40–49, April 2017.

[28] G.D. Finlayson, B. Schiele, and J.L. Crowley. Comprehensive colour image normalization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1406:475–490, 1998. ISBN: 9783540645696.

[29] Graham Finlayson, Steven Hordley, Gerald Schaefer, and Gui Yun Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38(2):179–190, February 2005.

[30] F. Garcia-Lamont, J. Cervantes, A. López, and L. Rodriguez. Segmentation of images by color features: A survey. *Neurocomputing*, 292:1–27, 2018.

[31] Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc., 2nd edition, September 2019.

[32] Rafael C. Gonzalez, Richard E. Woods, and Barry R. Masters. Digital Image Processing, Third Edition. *Journal of Biomedical Optics*, 14(2):029901, 2009.

[33] Paul D. Gordon, Courtney De Ville, James C. Sacchettini, and Gerard L. Coté. A portable brightfield and fluorescence microscope toward automated malarial parasitemia quantification in thin blood smears. *PLOS ONE*, 17(4):e0266441, April 2022. 2 citations (Crossref) [2023-12-06] 4 citations (Semantic Scholar/DOI) [2023-12-06] Publisher: Public Library of Science.

[34] Oscar Holmström, Sebastian Stenman, Antti Suutala, Hannu Moilanen, Hakan Kücükel, Billy Ngasala, Andreas Mårtensson, Lwidiko Mhamilawa, Berit Aydin-Schmidt, Mikael Lundin, Vinod Diwan, Nina Linder, and Johan Lundin. A novel deep learning-based point-of-care diagnostic method for detecting Plasmodium falciparum with fluorescence digital microscopy. *PLOS ONE*, 15(11):e0242355, November 2020. 5 citations (Crossref) [2023-12-06] Publisher: Public Library of Science.

[35] RW Horobin. How Romanowsky stains work and why they remain valuable — including a proposed universal Romanowsky staining mechanism and a rational troubleshooting scheme. *Biotechnic & Histochemistry*, 86(1):36–51, February 2011. 47 citations (Crossref) [2023-12-06] Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/10520295.2010.515491.

[36] A. Humeau-Heurtier. Texture feature extraction methods: A survey. *IEEE Access*, 7:8975–9000, 2019.

[37] N. Kanwal, F. Perez-Bueno, A. Schmidt, K. Engan, and R. Molina. The Devil is in the Details: Whole Slide Image Acquisition and Processing for Artifacts Detection, Color Variation, and Data Augmentation: A Review. *IEEE Access*, 10:58821–58844, 2022.

[38] Y.M. Kassim, F. Yang, H. Yu, R.J. Maude, and S. Jaeger. Diagnosing malaria patients with plasmodium falciparum and vivax using deep learning for thick smear images. *Diagnostics*, 11(11), 2021. 11 citations (Crossref) [2023-12-06].

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, January 2012.

[40] M.W. Lafarge, E.J. Bekkers, J.P.W. Pluim, R. Duits, and M. Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *Medical Image Analysis*, 68, 2021.

[41] Xintong Li, Chen Li, Md Mamunur Rahaman, Hongzan Sun, Xiaoqi Li, Jian Wu, Yudong Yao, and Marcin Grzegorzek. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artificial Intelligence Review*, 55(6):4809–4878, August 2022.

[42] T. Lindeberg. Image Matching Using Generalized Scale-Space Interest Points. *Journal of Mathematical Imaging and Vision*, 52(1):3–36, 2015.

[43] N. Linder, R. Turkki, M. Walliander, A. Mårtensson, V. Diwan, E. Rahtu, M. Pietikäinen, M. Lundin, and J. Lundin. A malaria diagnostic tool based on computer vision screening and visualization of Plasmodium falciparum candidate areas in digitized blood smears. *PLoS ONE*, 9(8), 2014. 79 citations (Crossref) [2023-12-06].

[44] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. Num Pages: 91-110 Place: New York, Netherlands Publisher: Springer Nature B.V.

[45] J. McVey and G. Finlayson. Least-squares optimal contrast limited histogram equalisation. volume 2019-October, pages 256–261, 2019. ISSN: 2166-9635.

[46] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[47] Kuo Niu and Chao Tian. Zernike polynomials and their applications. *Journal of Optics*, 24(12):123001, November 2022. Publisher: IOP Publishing.

[48] Ives Rey Otero. *Anatomy of the SIFT method*. phdthesis, École normale supérieure de Cachan - ENS Cachan, September 2015.

[49] O.O. Oyegoke, L. Maharaj, O.P. Akoniyon, I. Kwoji, A.T. Roux, T.S. Adewumi, R. Maharaj, B.T. Oyebola, M.A. Adeleke, and M. Okpeku. Malaria diagnostic methods with the elimination goal in view. *Parasitology Research*, 121(7):1867–1885, 2022. 15 citations (Crossref) [2023-12-06].

[50] Mahdieh Poostchi, Kamolrat Silamut, Richard J. Maude, Stefan Jaeger, and George Thoma. Image analysis and machine learning for detecting malaria. *Translational Research: The Journal of Laboratory and Clinical Medicine*, 194:36–55, April 2018. 272 citations (Crossref) [2023-12-06].

[51] CDC-Centers for Disease Control and Prevention. CDC - Malaria - Diagnosis & Treatment (United States) - Diagnosis (U.S.), January 2023.

[52] Sivaramakrishnan Rajaraman, Stefan Jaeger, and Sameer K. Antani. Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images. *PeerJ*, 7:e6977, May 2019. 91 citations (Crossref) [2023-12-06] Publisher: PeerJ Inc.

[53] R.T.C. Ramarolahy, E.O. Gyasi, and A. Crimi. Classification and Generation of Microscopy Images with Plasmodium Falciparum via Artificial Neural Networks Using Low Cost Settings. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12968 LNCS:147–157, 2021. 1 citations (Crossref) [2023-12-06] ISBN: 9783030877217.

[54] Sebastian Raschka. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, November 2020. arXiv:1811.12808 [cs, stat].

[55] Luís Rosado, José M. Correia da Costa, Dirk Elias, and Jaime S. Cardoso. Automated Detection of Malaria Parasites on Thick Blood Smears via Mobile Devices. *Procedia Computer Science*, 90:138–144, January 2016. 58 citations (Crossref) [2023-12-06].

[56] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432, 2015.

[57] U. Salamah, R. Sarno, A. Z. Arifin, A. S. Nugroho, M. Gunawan, V. Pragesjvara, E. Rozi, and P. B. S. Asih. Enhancement of low quality thick blood smear microscopic images of malaria patients using contrast and edge corrections. In *2016 International Conference on Knowledge Creation and Intelligent Computing (KCIC)*, pages 219–225, November 2016. 2 citations (Crossref) [2023-12-06].

[58] Sanghamitra Sathpathi, Akshaya K Mohanty, Parthasarathi Satpathi, Saroj K Mishra, Prativa K Behera, Goutam Patel, and Arjen M Dondorp. Comparing Leishman and Giemsa staining for the assessment of peripheral blood smear preparations in a malaria-endemic region in India. *Malaria Journal*, 13:512, December 2014. 31 citations (Crossref) [2023-12-06].

[59] M. Schinkel, K. Paranjape, R. S. Nannan Panday, N. Skyttberg, and P. W. B. Nanayakkara. Clinical applications of artificial intelligence in sepsis: A narrative review. *Computers in Biology and Medicine*, 115:103488, December 2019.

[60] S. Shambhu, D. Koundal, P. Das, V.T. Hoang, K. Tran-Trung, and H. Turabieh. Computational Methods for Automated Analysis of Malaria Parasite Using Blood Smear Images: Recent Advances. *Computational Intelligence and Neuroscience*, 2022, 2022. 6 citations (Crossref) [2023-12-06].

[61] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009.

[62] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer International Publishing, Cham, 2022.

[63] Amir Tahmasbi, Fatemeh Saki, and Shahriar B. Shokouhi. Classification of benign and malignant masses based on Zernike moments. *Computers in Biology and Medicine*, 41(8):726–735, August 2011.

[64] S.F. Tan and N.A.M. Isa. Exposure Based Multi-Histogram Equalization Contrast Enhancement for Non-Uniform Illumination Images. *IEEE Access*, 7:70842–70861, 2019.

[65] Michael Reed Teague. Image analysis via the general theory of moments*. *JOSA*, 70(8):920–930, August 1980. Publisher: Optica Publishing Group.

[66] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2007.

[67] M. Valkonen, K. Kartasalo, K. Liimatainen, M. Nykter, L. Latonen, and P. Ruusuvuori. Metastasis detection from whole slide images using local features and random forests. *Cytometry Part A*, 91(6):555–565, 2017.

[68] D. Vijayalakshmi, M.K. Nath, and O.P. Acharya. A Comprehensive Survey on Image Contrast Enhancement Techniques in Spatial Domain. *Sensing and Imaging*, 21(1), 2020.

[69] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, June 2014. Publisher: PeerJ Inc.

[70] Polrat Wilairatana, Noppadon Tangpukdee, and Srivicha Krudsood. Definition of hyperparasitemia in severe falciparum malaria should be updated. *Asian Pacific Journal of Tropical Biomedicine*, 3(7):586, July 2013.

[71] World Health Organization. Basic malaria microscopy. Technical report, World Health Organization, 2010. ISBN: 9789241547918 number-of-pages: 80.

[72] World Health Organization. Malaria microscopy quality assurance manual, December 2015.

[73] World Health Organization. *Microscopy for the detection, identification and quantification of malaria parasites on stained thick and thin blood films in research settings (version 1.0): procedure: methods manual*. World Health Organization, 2015.

[74] World Health Organization. World malaria report 2022, August 2022.

[75] Chaohong Wu, Joost Schulte, Katharine J. Sepp, J. Troy Littleton, and Pengyu Hong. Automatic Robust Neurite Detection and Morphological Analysis of Neuronal Cell Cultures in High-content Screening. *Neuroinformatics*, 8(2):83–100, June 2010.

[76] Feng Yang, Mahdieh Poostchi, Hang Yu, Zhou Zhou, Kamolrat Silamut, Jian Yu, Richard J. Maude, Stefan Jaeger, and Sameer Antani. Deep Learning for Smartphone-Based Malaria Parasite Detection in Thick Blood Smears. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1427–1438, May 2020. 103 citations (Crossref) [2023-12-06] Conference Name: IEEE Journal of Biomedical and Health Informatics.

[77] Feng Yang, Hang Yu, Kamolrat Silamut, Richard J. Maude, Stefan Jaeger, and Sameer Antani. Smartphone-Supported Malaria Diagnosis Based on Deep Learning. In Heung-Il Suk, Mingxia Liu, Pingkun Yan, and Chunfeng Lian, editors, *Machine Learning in Medical Imaging*, volume 11861, pages 73–80. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science.

[78] H. Yu, F.O. Mohammed, M. Abdel Hamid, F. Yang, Y.M. Kassim, A.O. Mohamed, R.J. Maude, X.C. Ding, E.D.A. Owusu, S. Yerlikaya, S. Dittrich, and S. Jaeger. Patient-level performance evaluation of a smartphone-based malaria diagnostic application. *Malaria Journal*, 22(1), 2023. 4 citations (Crossref) [2023-12-06].

[79] Hang Yu, Feng Yang, Sivaramakrishnan Rajaraman, Ilker Ersoy, Golnaz Moallem, Mahdieh Poostchi, Kannappan Palaniappan, Sameer Antani, Richard J. Maude, and Stefan Jaeger. Malaria Screener: a smartphone application for automated malaria screening. *BMC Infectious Diseases*, 20(1):825, November 2020. 24 citations (Crossref) [2023-12-06].

[80] Richard Zhang. Making Convolutional Networks Shift-Invariant Again. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7324–7334. PMLR, May 2019. ISSN: 2640-3498.

[81] Oliver S. Zhao, Nikhil Kolluri, Anagata Anand, Nicholas Chu, Ravali Bhavaraju, Aditya Ojha, Sandhya Tiku, Dat Nguyen, Ryan Chen, Adriane Morales, Deepti Valliappan, Juhi P. Patel, and Kevin Nguyen. Convolutional neural networks to automate the screening of malaria in low-resource countries. *PeerJ*, 8:e9674, August 2020. 16 citations (Crossref) [2023-12-06] Publisher: PeerJ Inc.

[82] Liang Zheng, Yi Yang, and Qi Tian. SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244, May 2018. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[83] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C.C. Loy. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023.

Glossary

List of Acronyms

AUC	Area Under Curve
CAD	Computer-aided diagnostics
CI	Confidence Interval
CNN	Convolutional Neural Network
DoG	Difference-of-Gaussian
FN	False Negative
FP	False Positive
FNR	False Negative Rate
HE	Histogram equalisation
IGMS	Iterative Global Minimum Screening
MAPE	Mean Absolute Percentage Error
PCR	Polymerase Chain Reaction
PRC	Precision-Recall Curve
AUPRC	Area Under Precision-Recall Curve
PSNR	Peak Signal to Noise Ratio
RDT	Rapid Diagnostic Tests
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SVM	Support Vector Matrix
TN	True Negative
TP	True Positive
WBC	White Blood Cell
WHO	World Health Organisation

List of Symbols

γ_b	Threshold: Minimum prediction score to label input as positive.
γ_f	Threshold to assign label to prediction, specifically for in-the-field performance.
D	Parasite density; Parasitaemia
$r_\top$	Annotated radius of parasite
d	Diameter