Biomedical Signal Processing and Control 94 (2024) 106289

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

A deep architecture based on attention mechanisms for effective end-to-end detection of early and mature malaria parasites

Luca Zedda *, Andrea Loddo *, Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124, Cagliari, Italy

ARTICLE INFO

Keywords:
Computer vision
Deep learning
Image processing
Malaria parasites detection
Early malaria diagnosis

ABSTRACT

Malaria is a severe infectious disease caused by the Plasmodium parasite. The early and accurate detection of this disease is crucial to reducing the number of deaths it causes. However, the current method of detecting malaria parasites involves manual examination of blood smears, which is a time-consuming and labor-intensive process, mainly performed by skilled hematologists, especially in underdeveloped countries. To address this problem, we have developed two deep learning-based systems, YOLO-SPAM and YOLO-SPAM++, which can detect the parasites responsible for malaria at an early stage. Our evaluation of these systems using two public datasets of malaria parasite images, MP-IDB and IML, shows that they outperform the current state-of-the-art, with more than 11M fewer parameters than the baseline YOLOv5m6. YOLO-SPAM++ demonstrated a substantial 10% improvement over YOLO-SPAM and up to 20% against the best-performing baseline in preliminary experiments conducted on the Plasmodium Falciparum species of MP-IDB. On the other hand, YOLO-SPAM showed slightly better results than YOLO-SPAM++ in subsets without tiny parasites, while YOLO-SPAM++ performed better in subsets with tiny parasites, with precision values up to 94%. Further cross-species generalization validations, conducted by merging training sets of various species within MP-IDB, showed that YOLO-SPAM++ consistently outperformed YOLOv5 and YOLO-SPAM across all species, emphasizing its superior performance and precision in detecting tiny parasites. These architectures can be integrated into computer-aided diagnosis systems to create more reliable and robust systems for the early detection of malaria.

1. Introduction

Malaria is a severe and potentially deadly disease caused by the Plasmodium parasite. This parasitic infection is spread primarily through the bites of female Anopheles mosquitoes infected with the parasite. In 2021, there were approximately 247 million malaria cases worldwide, with a staggering 619,000 deaths attributed to this disease. The majority of these cases (95%) and fatalities (96%) occurred in the World Health Organization (WHO) African region, with children under the age of five being the most vulnerable group, accounting for around 80% of deaths [1].

In humans, the parasites of the genus Plasmodium cause malaria by attacking red blood cells (RBCs), spreading to people through the bites of infected female Anopheles mosquitoes. Five species of the parasite can cause malaria in humans: P. falciparum (Pf), P. vivax (Pv), P. ovale (Po), P. malariae (Pm), and P. knowlesi (Pk), with Pf and Pv posing the most significant threat [1,2]. The life stages of the malaria parasite within the human host include the ring, the trophozoite, the schizont, and the gametocyte stages. Understanding these different stages is essential for developing effective treatments and prevention strategies for this dangerous disease.

The WHO defines human malaria as a preventable and treatable disease if diagnosed promptly, as the worsening illness can lead to disseminated intravascular thrombosis, tissue necrosis, and spleen hypertrophy [1,3]. Therefore, the key strategy is to diagnose the disease accurately and as early as possible and provide prompt treatment.

Malaria diagnosis can be accomplished using various diagnosis techniques, including microscopical analysis of blood smear, rapid diagnostic test (RDT), or real-time polymerase chain reaction (PCR). They can overcome the complications brought by the fact that symptoms can be easily confused with those of other diseases, such as viral hepatitis or dengue fever [4].

Despite its heightened accuracy, the PCR test is not ideal for program settings since it lacks the convenience of a point-of-care test compared to RDT or microscopy. Also, it requires specialized laboratory facilities to be conducted. To ensure accurate diagnosis in all situations, the WHO has recommended that all suspected malaria cases be verified

* Corresponding authors.
E-mail addresses: luca.zedda@unica.it (L. Zedda), andrea.loddo@unica.it (A. Loddo).

https://doi.org/10.1016/j.bspc.2024.106289
Received 23 September 2023; Received in revised form 4 January 2024; Accepted 25 March 2024
Available online 1 April 2024

through microscopy or an RDT before treatment. However, it is important to note that false-negative results can result in delayed treatment and an increased risk of spreading the disease.

Microscopy remains the method preferred by pathologists for diagnosing malaria [5–7], even in endemic countries, due to its sensitivity, affordability, and ability to identify parasite species and density [4,8,9]. It consists of analyzing a peripheral blood smear (PBS) on a glass slide to identify malaria parasites and their stages. This procedure is also widely used for other blood tasks, like leukemia detection [10–12] or blood cell counting [13,14].

However, microscopy has several drawbacks, as many issues can occur in this process: *(i)* detecting infections in the early stages can be challenging, and the presence of experienced microscopists is necessary; *(ii)* in some malaria-endemic regions, the scarcity of qualified microscopists, poor quality control, and misdiagnosis due to low parasitemia or mixed infections can limit microscopic diagnosis; *(iii)* some rural health facilities may not have access to this diagnostic method; *(iv)* identifying Plasmodium species can also be difficult under the microscope, leading to possible misreporting of certain species, such as P. ovale, which looks like P. vivax. Although this does not affect treatment because the patient will receive the same treatment for both species, it has crucial implications in the epidemiology and mapping of malaria [8,9]; *(v)* technical skills in slide preparation are required; *(vi)* lysis of red blood cells and related modifications in parasite morphology can happen, leading to errors in species identification; *(vii* the quality and illumination of the microscope are rarely guaranteed and are not standard; *(viii)* staining procedure can also affect the procedure; *(ix)* finally, the level of parasitemia plays a role as well [4].

Last but not least, there is the need to keep infectious diseases under control, especially in underdeveloped countries with no medical centers nearby or capable of handling many patients [3].

Accurate and timely malaria diagnosis is crucial for effective treatment and preventing severe complications. Although traditional methods like microscopy are still considered the gold standard, recent developments in deep learning have shown promising results in malaria cell image analysis, particularly with Convolutional Neural Networks (CNNs).

Several studies have explored the application of CNNs in malaria diagnosis at the single-cell level. These studies have highlighted the significance of accurately identifying whether a cell is infected with the malaria parasite [15–17].

However, using monocentric cell image datasets represents an overly ideal scenario in which salient and highly discriminating features can be extracted from the images. This scenario, more realistically, can be achieved by a previous step of detection or segmentation of a full-size image.

In real-world application scenarios, the systems are fully automated, and the images may not always be accurately centered or have perfect crops. This aspect can result in less-than-ideal detection and, thus, less precision in diagnosis. Previous studies have shown the effectiveness of detection systems in real-world application scenarios for computer-aided diagnosis (CAD) systems, including those that are robust and can handle such issues with image quality [18–22].

CAD systems can assist pathologists in diagnosing diseases and post-therapy monitoring. CAD systems excel at replicating manual analysis with significantly higher precision and faster results while minimizing subjectivity [23–25].

Additional challenges include distinguishing between various *Plasmodium* species and addressing the intricacies associated with low levels of parasitemia and asymptomatic infections. As a result, achieving precise bounding box detection for the exact localization of parasites within cells holds significant promise for comprehensive investigations and detailed diagnostic endeavors [7,26,27].

Therefore, incorporating deep learning techniques with object detection capabilities becomes imperative in this context, enabling the accurate classification of infected cells and the exact localization of parasites within them. This integration provides comprehensive information for detailed analysis and diagnosis.

The challenges and motivations presented so far motivated this work. Here, we present a new CAD approach for the automated, early identification of malaria parasites and early quantification of parasitemia with the dual purpose of assisting pathologists and overcoming the challenges described in gold-standard microscopy.

The main contributions are listed as follows. *(i)* We propose **YOLO-SPAM** and **YOLO-SPAM++**, two novel deep learning (DL)-based architectures projected for real-time, early detection of malaria through the identification of tiny parasites; *(ii)* the detection of four different malaria species and life stages, for mixed or intra-species detection is considered; *(iii)* an extensive evaluation of two different datasets and a comparison with three off-the-shelf object detectors is performed.

Our study employed the You Only Look Once (YOLO) architecture, a well-established approach that has demonstrated impressive results in our previous research [28]. Nevertheless, we implemented several architectural modifications to improve its accuracy in detecting tiny parasites to allow an early diagnosis and to perform multi-species detection of malaria parasites.

The remainder of this article is organized as follows. First, a background about the task at hand and an overview of the literature is given in Section 2. Then, materials and methods are described in Section 3, while Section 4 describes the proposed architectures. The experimental evaluation and the obtained results follow in Section 5. A detailed discussion and a description of the limitations are given in Section 6. Finally, conclusions are drawn in Section 7.

2. Background

CAD systems developed in medicine are not only limited to a particular area but can also be applied to hematology. Many CAD-based solutions have already been suggested for the automatic detection of malaria parasites in hematology. These solutions reduce manual analysis errors and offer a consistent interpretation of blood samples. Ultimately, this leads to a decrease in the cost of diagnosis [27,29].

By applying this promising technology to hematology, researchers hope to enhance the accuracy and speed of diagnosis in this field as well, thus leading to improved patient outcomes and a more efficient healthcare system [23].

Both traditional image processing methods and advanced deep learning techniques have been used in studies on automatic parasite detection. Conventional image processing involves detecting (or segmenting) the parasites, extracting features, and performing parasite classification that can be carried out independently or interrelatedly. In contrast, end-to-end deep learning approaches integrate all the steps and have become more prevalent after AlexNet's proposal [30].

Traditional pipeline methods in this area have included mathematical morphology techniques for preprocessing and segmentation [18,19]. Handcrafted feature extraction [21] has also been used to train machine learning classification methods.

Meanwhile, DL approaches have become more prominent over the last decade, with numerous works published [7,15,17,27,31].

These studies leverage DL techniques to improve the accuracy and efficiency of the detection process, often achieving superior performance compared to traditional methods.

The current literature on malaria parasite analysis from blood smear images can be organized into four main topics: *(i)* parasite detection and classification from full-size images (see Section 2.1; *(ii)* parasite classification from single-cell images (see Section 2.2; *(iii)* domain generalization methods from high- to low-cost devices (see Section 2.3); *(iv)* methods for low-cost sensor image devices (see Section 2.4).

2.1. Parasite detection and classification from full-size images

Detecting malaria parasites from blood images is a challenging task that requires analyzing full-size images depicting sections of a blood

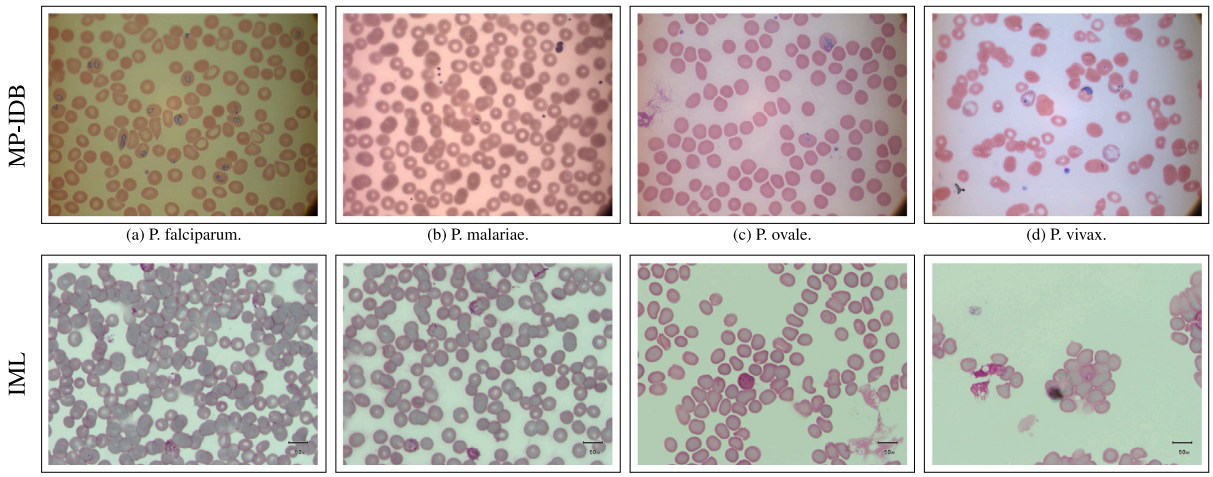


Fig. 1. Overview of the two datasets exploited in this study: MP-IDB and IML. MP-IDB contains four species of malaria: P. falciparum, P. malariae, P. ovale, and P. vivax. The IML dataset, on the other hand, only includes P. vivax samples. The MP-IDB dataset shows variations within each species, while the two datasets differ from each other.

smear. This analysis is critical for near real-time diagnosis, especially in settings where clinical facilities are limited, like underdeveloped countries where access to medical resources is restricted [21].

However, detecting parasites can be challenging because their structures can be similar to those of other cell regions, such as white blood cells and platelets. Microscopic evaluation of PBS typically takes more than 15 min per slide. Moreover, identifying the various species of Plasmodium can be challenging due to the parasites' different life-cycle stages [32].

Conducting a fine-grained examination is crucial to accurately diagnose diseases (e.g., malaria parasites, leukemia) from images of blood samples. This kind of examination requires segmentation or detection techniques to identify regions of interest (ROIs) before performing a further classification and providing the pipeline with relevant information for a thorough understanding of these regions [33].

However, some works involve directly classifying full-size images, often utilizing CNNs or traditional machine learning techniques, trained using handcrafted features or those extracted from pre-trained CNNs. An example of this strategy is the work of Vijayalakshmi et al. [5], who used an SVM trained with features extracted from a VGG-19 net to distinguish between infected and non-infected malaria images.

In this context, the newest approach for handling a multi-stage pipeline involves deep learning techniques. For example, Arshad et al. [31] proposed a method based on a segmentation step that combines U-Net and watershed algorithms, followed by a binary classification to separate healthy and infected cells and a further life cycle-stage classification for the infected cells. Both classifications are based on ResNet50v2's CNN.

Similarly, Maity et al. [7] adopted a semantic segmentation followed by a Capsule Network to classify Pf rings. On the contrary, instead of using a segmentation approach, Sultani et al. [27] performed a comparison of different off-the-shelf object detectors, viz. Faster R-CNN (FRCNN), RetinaNet, FCOS, and YOLO on Pv's life stages. In contrast, Lin et al. [34] and Manescu et al. [35] proposed two custom object detection (OD) pipelines to diagnose malaria and identify the parasites.

Our research has identified three other studies that also use YOLO to detect parasites, but they focus on thick blood smear images and use datasets different from ours. They used YOLO versions, including YOLOv3 and YOLOv4 [36–38].

2.2. Parasite classification from single-cell images

Since malaria parasites always affect the RBCs, these methods focus on the classification of individual cells from images to distinguish between parasitized and healthy erythrocytes [15–17,39,40].

These studies are typically benchmarked on the NIH dataset. More recently, there have been investigations into using vision transformers on the same dataset. These investigations, as explored in Sengar et al. [41] with a classification of Pv life stages, are part of a growing trend that seeks to explore the potential benefits of vision transformers in deep learning-based malaria research.

The specific solutions in this sense are novel ad-hoc designed CNN architectures [15], use of the transfer learning on CNNs pre-trained on ImageNet [42], e.g., ResNet-50 [16] and VGG-19 [39], or ensemble with VGG-19 and SqueezeNet [17]. In a recent study, Diker et al. [40] presented a residual CNN architecture that uses Bayesian optimization to extract key features from both classes. The identified features are then used as input to an SVM classifier.

2.3. Domain generalization methods from high- to low-cost devices

When it comes to computer-aided medical image analysis, machine learning techniques often encounter a challenge known as the domain shift problem caused by different distributions between source data and target data. Domain adaptation has emerged as a potential solution and has gained significant attention in recent years [43].

In their study, Sultani et al. [27] tackled the challenge of obtaining images in areas with limited medical resources by collecting a dataset using low-cost and high-cost microscopes. They tested different domain adaptation techniques to determine the most suitable way to use high-cost microscope images as the source domain and low-cost microscope images as the target domain.

Further possible domain adaptation tasks still need to be addressed in this field. One example is the ability to classify different Plasmodium species based on knowledge of only one species (e.g., being able to recognize Pm, Pv, and Po when only Pf is the source domain).

2.4. Methods for low-cost sensor image devices

Low-cost mobile devices like smartphones and tablets, often paired with microscope cameras, have been used in studies for image acquisition and analysis processes. Smartphone-specific apps typically based on pre-trained or customized CNN have been developed to automate malaria diagnosis [21,44], resulting in excellent classification rates in as little as ten seconds [44].

The use of affordable and easy-to-use mobile devices has significantly expanded in recent years, especially in resource-limited countries with a high incidence of malaria deaths and a lack of specialized personnel and equipment for proper diagnosis [44]. This technology can provide a cost-effective solution for accurate malaria diagnosis [32].

2.5. Limitations of the existing literature

There are some limitations in the existing literature regarding analyzing full-size images. Direct classification using CNN or traditional machine learning techniques may oversimplify the task, leading to the loss of fine-grained details essential for accurate diagnosis. Therefore, some studies opt for off-the-shelf object detectors on custom datasets. However, previous works on malaria parasite detection mainly focused on thick blood smears. They may not be generalized to other datasets, as differences in datasets and features can significantly affect model performance. Furthermore, there is a lack in the analysis of multiple malaria species and life stages since the existing literature's main goal is to address a specific species.

Our work has focused on analyzing two public thin blood smear image datasets. This choice has allowed us to perform a detailed examination of the fine-grained details in the images. Such an analysis helps in the identification and detection of different species and stages of life, both from a quantitative and qualitative perspective.

Additionally, as depicted in Fig. 1, the two datasets exhibit distinct intrinsic characteristics, including variations in coloration, illumination conditions, composition, and types of parasites. This diversity enables the demonstration of the proposed method's capability to not only tackle the detection of tiny parasites but also address the identification of fully developed parasites within the context of an infection.

3. Materials and methods

This section describes the datasets employed in Section 3.1 and gives an overview of the state-of-the-art object detectors in Section 3.2 and the YOLO family in Section 3.3. Furthermore, it delves into the concept of attention and its applications in the field (Section 3.4). Finally, insights on Swin transformers are provided in Section 3.5.

3.1. Datasets

Some samples of the two public datasets used in this work are shown in Fig. 1.

MP-IDB is an image dataset including 210 pictures of four types of malaria species. These are made up of 104 *P. falciparum*, 37 *P. malariae*, 29 *P. ovale*, and 40 *P. vivax* images. The life cycle of every species comprises four distinct stages: *ring*, *trophozoite*, *schizont*, and *gametocyte* [6].

Each picture has a corresponding ground truth that indicates the presence of one or more of these life stages. The images were taken at a high resolution of 2592 × 1944 pixels and with a 24-bit color depth. As shown in Fig. 1, the dataset features significant variation within and across species.

IML [45] comprises images of blood samples collected from individuals infected with malaria in Pakistan's Punjab province. The images were taken with a camera attached to a microscope from the XSZ-107 series, magnified at 100×. The dataset comprises 345 images, each containing an average of 111 blood cells. The only malaria species represented is *P. vivax*. Each image has its corresponding ground truth that indicates one or more life stages or red blood cells. The images have a resolution of 1280 × 960 pixels with a 24-bit color depth.

3.2. Object detectors

Object detection is a critical task based on deep learning object detectors in computer vision. They are conventionally divided into two categories: two-stage and one-stage.

Two-stage architectures, such as Faster R-CNN [46], first identify ROIs and then perform classification and bounding box regression in a coarse-to-fine process. In contrast, one-stage detectors, including RetinaNet [47], FCOS [48], and YOLO family [49], produce bounding boxes and classes directly from predicted feature maps with predefined anchors.

Two-stage architectures generally provide slightly higher accuracy, while single-stage detectors are faster and more compact, making them more suitable for time-critical applications and computationally constrained edge devices [47,50,51].

More recently, the success of Transformers in image recognition has led to the development of Swin Transformers-based (see Section 3.5), or end-to-end DEtection TRansformers (DETRs). Despite their high recognition accuracy, DETRs are hampered by their complex architectures and slow convergence problems [50].

To address the existing limitations, we propose a modified version of the one-stage YOLOv5 detector's architecture to enhance its accuracy and efficiency in detecting malaria parasites, particularly the smallest and earliest ones.

3.3. The YOLO family of detectors

The YOLO family of detectors uses a different approach compared to traditional methods. Instead of a two-step process based on region selection, it uses an end-to-end differentiable network that integrates bounding box estimation and object identification. The input image is divided into $S \times S$ constant-size grids, and a CNN predicts bounding boxes and classes for each grid. If the confidence of a bounding box is higher than a fixed threshold, it is selected to locate the object in the image. The CNN performs one pass and produces known objects and their bounding boxes, ensuring that each object is detected only once after non-maximum suppression.

However, despite their significant improvement in detection speed, YOLO architectures struggle to detect small objects compared to two-stage detectors [49,51]. This limitation was considered one of the objectives of the proposed work, as the scenario includes cases in which the first parasites appear tiny. The smallest ones, i.e., the smallest rings, are typically not large enough to be considered by a generic detector.

YOLOv5. This architecture has been chosen as a baseline for our purpose because of its speed, accuracy, and ease of training.

It is a family of OD architectures and models pre-trained on the Common Object in Context (COCO) dataset [52] and used for various object detection tasks [53].

The family includes five different models that share the same architecture but differ in size and complexity: *YOLOv5n* for nano, *YOLOv5s* for small, *YOLOv5m* for medium, *YOLOv5l* for large, and *YOLOv5x* for extra-large models.

Each model is available pre-trained on 640 × 640 or 1280 × 1280 resolution images, with varying numbers of trainable parameters. For the first dimension, *YOLOv5n* contains 1.9 million parameters, *YOLOv5s* 7.2, *YOLOv5m* 21.2, *YOLOv5l* 46.2 and *YOLOv5x* 86.7. The latter contains 3.2, 12.6, 35.7, 76.8, and 140.7 million parameters, respectively.

The architecture of YOLOv5 consists of three components, similar to other object detection models: backbone, neck, and prediction head. The backbone is a pre-trained network dedicated to image feature extraction, and the neck combines the extracted features and creates three different scales of feature maps (also known as feature pyramids) to help the model generalize well to objects of different sizes and scales. The prediction head applies anchor boxes to the feature maps and detects objects based on the previously created feature maps. YOLOv5 uses the CSPDarknet53 architecture with a Spatial Pyramid Pooling (SPP) layer [54] as the backbone, Path Aggregation Network (PANet [55] as the neck, and the YOLO detection head [49].

3.4. Attention mechanism

Attention is a crucial cognitive process that affects how humans perceive the world. Instead of processing an entire visual scene simultaneously, humans selectively focus on the most important parts to capture its structure more accurately [56]. This mechanism helps

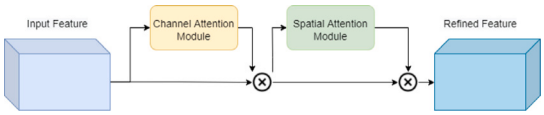


Fig. 2. Schematic representation of the architecture of the Convolutional Block Attention Module. Two consecutive attention sub-modules refine feature maps in channel and space, respectively.

filter out unnecessary information, making perceptual processing more efficient and accurate [57,58].

Recently, attention has become increasingly relevant in the computer vision field [57,58]. Here, attention focuses on specific input data when generating an output. The process involves weighting the importance of different input features to produce a set of weights for each feature, followed by a weighted sum to generate the output.

The attention module's structure involves two sets of vectors: $x1$ and $x2$. $x2$ generates a 'query,' while $x1$ creates a 'key' and 'value' pair. The attention function aims to connect the query with the key–value pairs to produce an output. This output is achieved by calculating a weighted sum of the value vectors. The compatibility function assigns weights based on the similarity between the query and each key. The output is obtained by taking the dot product between the softmax of the compatibility scores and the values, as discussed in [59].

In formal terms, when provided with a group of input features labeled as $x_1, x_2, \ldots, x_n$ and a desired output labeled as y, the attention mechanism calculates a weighted sum of the input features using the formula shown in Eq. (1) [59].

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

Various attention mechanisms are available, but two of the most commonly used are *spatial attention*, and *channel attention* [57,58].

Spatial attention focuses on identifying the important positions in an image that the neural network needs to learn. This mechanism transforms the spatial information of the original picture into another space while retaining the key information.

Channel attention focuses on the inter-channel relationship of features to create a channel attention map. This mechanism considers each channel in a feature map as a feature detector and concentrates on what is meaningful in the input image.

A combination of the two mechanisms, namely *Convolutional Block Attention Module (CBAM)*, was proposed by Woo et al. to improve informative channels and significant regions [56]. In particular, CBAM uses two sub-modules, channel and spatial. It separates the channel and spatial attention maps to make computations more efficient. It also introduces global pooling to utilize global spatial information.

Combining channel and spatial attention in sequence helps the network understand the relationships between different features. This way, the network knows what to focus on and where to focus.

Attention mechanisms play a critical role in enhancing the performance of object detection models. These mechanisms work by refining feature maps and are typically included in the model. In OD tasks, attention modules are designed to focus on three-dimensional feature maps and learn both channel-related features and spatial attention. However, the CBAM module uniquely separates channel and spatial awareness into two distinct sub-modules. A schematic representation of CBAM is shown in Fig. 2.

NAM.A modified version of the CBAM module created for image classification is the *Normalized Attention Module (NAM)*. It was designed to address the common problem of varying dot-product attention scores, which can be influenced by the input's dimensionality [60]. This inconsistency can lead to training instability and affect the attention mechanism's quality. The NAM solves this issue by normalizing the dot-product attention scores and dividing them by the square root of the input's dimensionality. It helps to stabilize the attention mechanism during training and ensures that the scores are appropriately scaled, regardless of the input dimensionality, providing the network with more stability. The normalization occurs before the softmax function is applied to calculate the attention weights.

3.5. Swin transformers

Microsoft Research introduced the Swin Transformer in 2021 as a new application for computer vision tasks [61]. This transformer-based approach uses multi-headed self-attention modules to process patches of input images that are converted into embeddings. The Swin Transformer allows for linear computation complexity with image size and enables cross-window connection, resulting in more accurate detection. Although it requires more parameters than convolutional models, Swin Transformer can replace convolution for vision tasks.

This model uses hierarchical feature maps, like those in CNN, that down-sample images by 4×, 8×, and 16×. This backbone helps with tasks such as object detection and instance segmentation.

In this work, we explored using Swin Transformer to gather global information. To improve the detection of tiny parasites, we enhanced the detection heads of the YOLOv5 model by adding a target detection head. Our modification included the integration of C3STR layers into the original C3-based structure of the YOLOv5 model, which enhanced its capability to gather feature information.

4. Proposed approach

This section provides an overview of our customizations to the YOLO architecture proposed to solve the problem faced. We will first introduce the architectural concept of YOLO-SPAM and its details in Section 4.1. Following that, in Section 4.2, we will describe the two proposed architectures and their relative additions and implementations of attention blocks.

4.1. The proposed networks: YOLO-SPAM and YOLO-SPAM++

This work aims to create an accurate malaria parasite detector by incorporating attention modules to improve the current methods of detecting malaria parasites by addressing their inherent limitations.

Our goals are to: *(i)* obtain the speed and compactness of one-stage detectors but achieve high accuracy without the need for a subsequent phase; *(ii)* identify small parasites within the same system; *(iii)* integrate Transformers without over-complicating or slowing down the architecture.

The YOLOv5 model was chosen as the baseline for its convenience and efficiency as a one-stage detector [53]. Specifically, we used the *YOLOv5m6* version pre-trained on 1280 × 1280 pixels images, which has 41.2 million trainable parameters.

This model balances network depth and parameter count, making it suitable for use on low-end machines and mobile devices with limitations for real-time detection of malaria parasites [6,21].

We propose two architectures: a base version, YOLO-SPAM, and its relative extension, YOLO-SPAM++. Both architectures use CBAM modules, but YOLO-SPAM++ includes NAM and Swin Transformer modules and a feature merging strategy.

The key idea behind our additional layers is related to the lack of inherent attention mechanisms in the YOLOv5 architecture. Attention mechanisms aim to refine feature maps from intermediate layers to improve detection results while minimizing computational overhead. Specifically, we introduced the CBAM, composed of spatial and channel attention modules, as defined in Section 3.4. Prior research has shown that CBAM is a practical addition and improves classification and detection tasks [56].

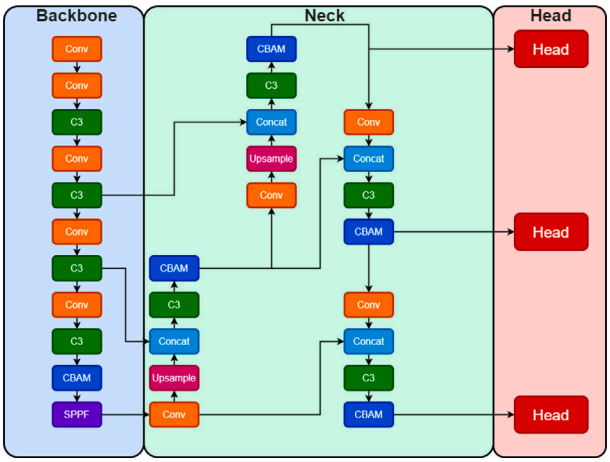


Fig. 3. The YOLO-SPAM architecture is presented, incorporating Convolutional Block Attention Modules (CBAM) to enhance the detection of the smallest parasites. These attention modules are strategically integrated into the backbone and neck, contributing to the effective identification of malaria parasites, especially those of small size. The additional CBAM layers are depicted in blue.

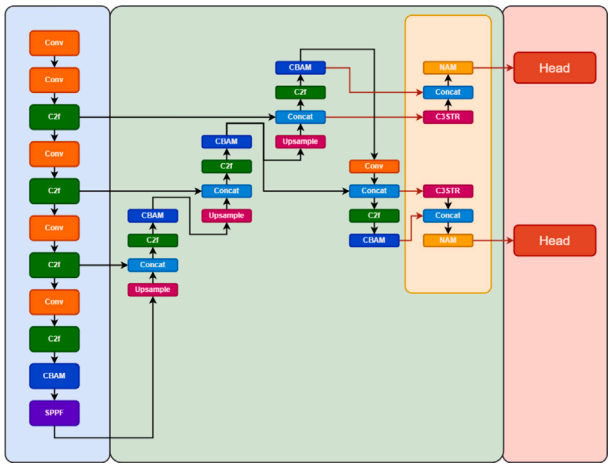


Fig. 4. YOLO-SPAM++ architecture featuring Convolutional Block Attention Modules (CBAM), Normalized Attention Modules (NAM), and Swin Transformers' C3STR modules. This model has two prediction heads, significantly enhancing its capability for detecting small objects. The feature arcs connected to the novel feature merging strategy are highlighted with red arrows, and the layers related to the merging strategy are emphasized in an orange box.

We studied where to insert such layers to maximize benefit while requiring the fewest additions. In both architectures, we strategically place a CBAM layer immediately after the last C3 layer in the backbone and near each prediction head. This choice is based on the shape of the features within the selected layers. Deeper feature maps have reduced spatial resolution but higher-dimensional channel information. By leveraging the CBAM module, we prioritize channel-related attention, which is computationally less demanding using 1×1 convolutions. In contrast, spatial attention requires a 7×7 convolution. Not only does this placement reduce computational requirements, but it also allows for a larger receptive field for spatial attention due to the limited spatial dimension.

Both architectures are described as follows.

4.2. YOLO-SPAM proposed architectures

YOLO-SPAM In Fig. 3, YOLO-SPAM is presented. As can be seen from the figure, our research introduces multiple attention modules integrated into the backbone and neck of the model, along with a new prediction head specifically designed to detect tiny object parasites. This innovative head utilizes information from the second neck block, which acts as a lower-level feature map extractor, achieved by concatenating the data from the second backbone C3 layer right before the head's C3 layer.

This novel approach enhances predictions across various object sizes and serves as the primary means for detecting small objects in the network. Higher-level feature maps can sometimes collapse the intensity of parasite presence into small, point-like spaces due to reduced spatial dimensions. For example, multiple instances of ring-stage parasites within a single red blood cell may be treated as a single prediction, leading to reduced spatial resolution. Similarly, smaller spatial resolutions may overlook a single-ring stage with a low pixel count.

However, lower-level feature maps maintain higher spatial resolution, mitigating the risk of prediction collapse. By leveraging these maps and incorporating attention modules, our research addresses these challenges. It significantly improves the model's ability to accurately detect and classify malaria parasites, particularly those that are tiny in size.

YOLO-SPAM++. The architectural design of this model is depicted in Fig. 4. It incorporates NAM and C3's Swin Transformer modules into the YOLO-SPAM architecture, resulting in a more advanced and lighter design.

To enhance the performance of YOLO-SPAM, the refined features from the CBAM layers are merged with those obtained from sequential C3STR layers (see Section 3.5). In this way, two new models specialized in detecting small objects are proposed. Also, a NAM module was then used to assign higher weight to the most important features and lower weight to less significant ones, further enriching the merged features.

The lower backbone layers extract less refined features but have a higher resolution, allowing them to detect parasites with minimal pixels. On the other hand, the higher levels can more accurately distinguish medium-sized parasites while potentially losing information about small parasites due to the lower feature map resolution obtained through the convolution process.

Compared with the YOLO-SPAM, the YOLO-SPAM++ architecture has fewer parameters and two prediction heads than three, making it suitable for small and medium-sized objects.

Research has shown that when transformers and attention are combined, they can enhance the YOLOv5 framework. By using vision transformers to extract global features, these techniques improve the model's performance compared to its original version. One way to achieve this is by replacing the last C3 layer with transformer blocks [53,62]. However, our approach takes a different direction. We aim to integrate local specialized features of the CNN with the global features in the model heads. To do this, we added C3STR modules while retaining the original C3 ones. In simpler terms, our approach maintains both local and global features in the model.

5. Experimental results

The primary objective of the experiments is to evaluate the effectiveness of the proposed architectures in improving the detection of malaria parasites, with a particular focus on identifying tiny ring-stage parasites. By achieving early detection, we can take proactive measures and combat the disease more effectively.

This section begins by outlining the experimental setup (Section 5.1) and continues with a preliminary experiment conducted on the Pf split of MP-IDB (Section 5.2). Then, the description of the experimental results obtained on MP-IDB and IML are given in Sections 5.3 and 5.4, respectively. The performance of the proposed architectures is further evaluated with two cross-dataset experiments, described in Section 5.5. Additionally, a qualitative analysis is provided in Section 5.6.

Table 1
Composition of the two datasets considered in the experimental evaluation. The table shows the experimental splits adopted for the four species of parasites for MP-IDB, and for IML.

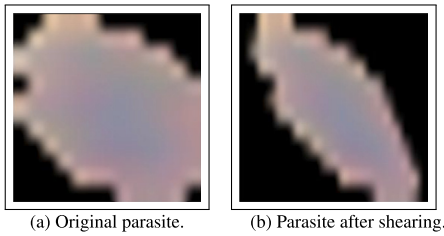
Dataset	Species	Train set imgs		Val set imgs	Test set imgs
		or. size	aug. size	or. size	or. size
MP-IDB	P. falciparum	83	2,905	10	11
	P. malariae	29	1,015	4	4
	P. ovale	23	805	3	3
	P. vivax	32	1,120	4	4
IML	P. vivax	241	8,435	35	69

Table 2
Distribution of the parasites of both datasets based on their size, measured in pixels. S, M and L indicate small, medium and large parasites.

Dataset	Species	Parasites								
		Train set			Val set			Test set		
		S	M	L	S	M	L	S	M	L
MP-IDB	P. falciparum	370	408	0	123	136	0	123	136	1
	P. malariae	1	25	0	1	8	0	0	8	0
	P. ovale	0	20	0	0	6	0	0	7	0
	P. vivax	2	29	2	1	10	1	3	10	5
IML	P. Vivax	6	128	249	1	9	49	3	16	89

Table 3
Image augmentation parameters for models training.

Augmentation	Parameters	Probability
Rotation	range iterations: [0, 3]	1
Gaussian Noise	variance range: [50, 100]	0.3
HSV - Hue	shift limit: 20	0.3
HSV - Saturation	shift limit: 30	0.3
HSV - Value	shift limit: 20	0.3



(a) Original parasite. (b) Parasite after shearing.

Fig. 5. This image illustrates how incorrect augmentations can distort small parasites, making them appear different from their actual shape and becoming unrepresentative of their class. In this example, shearing augmentation is applied. The ring stage, represented by a circular pattern with a pixel intensity spike, is compromised by the shear transformation. Another important issue is related to the pixel count for smaller instances of this class; in the image, the original ring stage parasite has halved its width, making the detection task more challenging.

5.1. Setup

The experiments were conducted on a workstation equipped with an Intel(R) Xeon(R) Gold 6136 CPU @ 3.00 GHz CPU, 64 GB RAM, and an NVIDIA Tesla P6 GPU with 16 GB memory.

We utilized the PyTorch implementation of **YOLOv5**,[1] created by the Ultralytics LLC team [63], and the yoloair's implementation of **C3STR**.[2] In addition, **Faster R-CNN**, **RetinaNet**, and **FCOS** were trained using the Detectron library [64].

Every YOLO-based architecture was initialized with pre-trained weights from the **COCO2017** dataset [52]. For YOLO, Darknet53 was selected as the backbone, while the other detectors utilized ResNet-50 pre-trained on ImageNet.

Every method was trained with the following hyperparameters: Adam was set as the optimizer with a weight decay of 1×10^{-2} and momentum of 0.9. The initial learning rate was 1×10^{-4} across a total of 100 epochs. Dropout was set to 0.2.

Datasets split. The MP-IDB dataset was divided into three splits: training, validation, and testing, with each parasite class divided into 60% for training, 20% for validation, and 20% for testing. The splits were constructed for IML using the guidelines provided by the authors [31]. Details are given in Table 1. Additionally, Table 2 furnishes the specific details on the dimensions of the parasites.

Data augmentation. To improve the accuracy of our models and address the issue of data imbalance, we created 35 augmented samples for each species based on the original data. This approach also helps to enhance the models' ability to handle object rotations and generalize to different scenarios. However, we used a lighter augmentation method to avoid damaging small parasites, identified as a significant concern in a previous study (see [28]). We avoided techniques such as shearing that could potentially harm these small parasites, as illustrated in Fig. 5. Table 3 details the augmentation methods used.

Metrics. Object detection methods are commonly evaluated with mean *average precision* (AP) metric and its variants [65]. Precision uses the Intersection over Union (IoU) concept to determine detection accuracy. Specifically, the IoU is the ratio of the overlap area between the predicted bounding box and the actual object compared to the total area of both. If the IoU is above a certain threshold, the detection is correct and labeled as a *true positive* (TP). However, if the IoU falls below the threshold, the detection is considered a *false positive* (FP). Additionally, if the model fails to detect an object present in the ground truth, this is referred to as a *false negative* (FN).

$$P = \frac{TP}{TP + FP} \tag{2}$$

where:

- *true positives (TP)* represents the number of instances belonging to the *positive* class that have been correctly predicted;
- *false positives (FP)* indicates when a nonexistent object is incorrectly detected or an existing object is detected in the wrong location;
- *false negatives (FN)* represents when a ground truth bounding box goes undetected.

In general, *precision* is defined in Eq. (2). In this work, the experimental evaluations were conducted considering five variants of the mAP metric:

[1] Available at the official repository.
[2] Available at the official repository.

Table 4
Results of the preliminary study on the Pf subset of MP-IDB. The table shows the results of YOLO-SPAM and YOLO-SPAM++ against the three variants of the YOLOv5 baseline architecture (small, medium and large).

Method	AP (%)	AP_{50}	AP_s	AP_m	AP_L
YOLOv5s6	57.5	92.6	55.7	59.2	0.0
YOLOv5m6	62.5	93.5	57.8	66.4	**100.0**
YOLOv5l6	63.6	93.6	56.6	69.2	**100.0**
YOLO-SPAM	74.7	98.7	65.6	69.8	**100.0**
YOLO-SPAM++	**84.6**	**99.1**	**75.5**	**80.3**	**100.0**

- **AP** is evaluated with 10 different IOUs varying in a range of 50% to 95% with steps of 5%;
- **AP_{50}** is evaluated with a single values of IOU corresponding to 50%;
- **AP_s** is the AP determined for small objects (with area $< 32^2$ pixels);
- **AP_m** is the AP determined for medium objects (with $32^2 <$ area $< 96^2$ pixels);
- **AP_L** is the AP determined for large objects (with area $> 96^2$ pixels).

5.2. Preliminary study on P. falciparum

We conducted a preliminary study using the two proposed architectures against the YOLOv5 architecture to compare their performance on the P. falciparum species of MP-IDB. This choice has been made for the following reasons: *(i)* Pf is the most numerous species in MP-IDB; *(ii)* Pf contains an adequate number of small, medium, and large parasites, as shown in Table 2.

We trained each architecture for 100 epochs to ensure fairness using the same structures, hyperparameters, and settings. After training, we evaluated the models using identical hyperparameters and settings for inference, leading to comparable results.

This experiment's results, shown in Table 4, indicate that both YOLO-SPAM and YOLO-SPAM++ significantly increased AP by 12.2% and 22.1%, respectively. The improvement is satisfactory even concerning the YOLOv5l6 baseline model, by 11.1% and 21%, respectively. This performance suggests that the addition of CBAM and the adoption of C3STR and NAM are effective when integrated into the model.

5.3. Experimental results on MP-IDB

Species-specific detection. Table 5 shows the performance obtained by both YOLO-SPAM architectures against four state-of-the-art object detectors (i.e., Faster R-CNN, RetinaNet, FCOS, and YOLOv5m6) on the four parasite species included in MP-IDB. Some performance values are missing because that particular sample was not included in the subset considered, e.g., Pm and Po do not contain any small or large parasite in their test sets (see Table 2).

The results obtained on P. falciparum show that YOLO-SPAM++ outperforms all the other methods in all reported metrics. YOLOv5m6 also performs well on this subset, with an AP of 62.5%. Interestingly, the FCOS method performs the worst with an AP of only 10.1%. In addition, this subset contains an extensive presence of tiny parasites, given the distribution shown in Table 2. The performance of YOLO-SPAM++ and YOLO-SPAM are the best even in this case, as they reach 75.5% and 65.6% of AP_s respectively, about 20% and 8% above YOLOv5m6, which is the third best detector. These results show the effectiveness of the proposed architectures in detecting tiny parasites.

As for P. malariae, the best performing method is YOLO-SPAM with an AP of 94.1% and AP_{50} of 99.5%. YOLOv5m6 and YOLO-SPAM++ also perform well, with AP of 80.0% and 93.6%, respectively. FCOS performs the worst again, with an AP of only 4.7%.

As for P. ovale, YOLO-SPAM is the best-performing method with an AP of 93.8% and AP_{50} of 99.5%. YOLO-SPAM++ and YOLOv5m6 also perform well, with AP of 87.4% and 83.9%, respectively.

The most effective method for detecting P. vivax is YOLO-SPAM++, which has an AP of 87.5%. YOLO-SPAM is a close second, with an AP of 83.6%. It should be noted that all methods struggle to detect small parasite objects. Our proposed architectures can achieve a 15.2% detection rate, making them the third-best performer after YOLOv5m6 and FRCNN. In this subset, the low results are considered acceptable because the number of training samples for tiny Pv parasites is meager, as indicated in Table 2.

Overall, it can be observed that YOLO-SPAM-based methods perform better than the other four methods on all four subsets. YOLO-SPAM++ performs the best among all the methods, achieving the highest AP for Pf and Po and the second-highest AP for Pm and Pv. YOLO-SPAM also performs well, especially on Pm and Po subsets. FCOS performs the worst among all the methods on all four datasets. In conclusion, the results suggest that YOLO-SPAM-based methods, especially YOLO-SPAM++, are more effective than the other methods reported for malaria parasite detection.

Species-generalization detection. In order to address the possibility of detecting mixed or intra-species, we propose an approach that involves training both models with the complete training set of all four species simultaneously. More precisely, in Table 7, it is shown that YOLO-SPAM generally outperforms YOLOv5m6 for all species in terms of AP, with the highest improvement observed for P. falciparum and P. vivax. For P. falciparum, YOLO-SPAM achieved an AP of 68.1% compared to 62.9% for YOLOv5m6, while for P. vivax, the AP values are 73.1% and 68.9% for YOLO-SPAM and YOLOv5m6, respectively. However, both methods show similar performance for P. malariae and P. ovale, with YOLOv5m6 achieving slightly higher results for both species.

Overall, the results suggest that the proposed YOLO-SPAM method can improve the detection performance for malaria parasites in a mixed or intra-species scenario, even for tiny parasites, compared to the baseline YOLOv5m6 method.

5.4. Experimental results on IML

Table 8 presents the performance obtained by our two proposed architectures on the IML dataset [31].

As can be seen, they outperform the baseline established with YOLOv5m6. In particular, YOLO-SPAM++ obtained an AP of 61.5%, while YOLO-SPAM achieved the best overall performance with an AP of 62.0%, outperforming YOLO-SPAM++ on every other metric. These results suggest that YOLO-SPAM-based architectures can be considered effective object detection methods even for the IML dataset compared to the baseline, as they both improve it.

To the best of our knowledge, only the dataset's authors report the detection performance on this dataset [31]. In particular, they disclosed their findings on the detection of both healthy and infected red blood cells by employing two segmentation approaches. They reported that the morphological method had an 89.3% bounding box precision, while the U-net-based approach had an 82.4% precision rate. Notably, these results included healthy cells, which makes it unfair to compare them directly to our approaches that solely target infected cells. However, we would like to highlight that in addition to the results presented in Table 8, our YOLO-SPAM method achieved a bounding box precision of 80.5%, while YOLO-SPAM++ achieved 89.0%.

5.5. Cross-dataset results

The last evaluation of the proposed architectures is devoted to establishing the performance with two cross-dataset experiments, reported in Table 9. In particular, considering the models trained on the Pv

Table 5
Quantitative evaluation on the four classes of parasites present in MP-IDB. Best results are emphasized in bold.

Class	Method	AP (%)	AP_{50}	AP_s	AP_m	AP_L
P. falciparum	FRCNN	39.2	80.6	33.7	44.3	0.0
	RetinaNet	34.0	78.5	23.9	42.6	0.0
	FCOS	10.1	39.9	5.6	14.5	0.0
	YOLOv5m6	62.5	93.5	57.8	66.4	**100.0**
	YOLO-SPAM	74.7	98.7	65.6	69.8	**100.0**
	YOLO-SPAM++	**84.6**	**99.1**	**75.5**	**80.3**	**100.0**
P. malariae	FRCNN	75.1	98.4	–	75.1	–
	RetinaNet	76.0	95.0	–	76.2	–
	FCOS	4.7	21.2	–	8.8	–
	YOLOv5m6	80.0	96.4	–	72.0	–
	YOLO-SPAM	**94.1**	**99.5**	–	**85.9**	–
	YOLO-SPAM++	93.6	98.5	–	84.2	–
P. ovale	FRCNN	71.0	89.1	–	71.0	–
	RetinaNet	74.3	91.5	–	74.3	–
	FCOS	44.2	81.8	–	45.1	–
	YOLOv5m6	83.9	96.8	–	76.3	–
	YOLO-SPAM	**93.8**	**99.5**	–	**83.9**	–
	YOLO-SPAM++	87.4	92.8	–	79.2	–
P. vivax	FRCNN	60.3	87.7	20.2	61.5	85.0
	RetinaNet	62.8	85.5	10.1	65.7	84.1
	FCOS	53.0	81.0	5.1	53.8	83.1
	YOLOv5m6	83.1	93.2	**21.9**	79.8	92.5
	YOLO-SPAM	83.6	92.9	15.2	79.2	89.8
	YOLO-SPAM++	**87.5**	**93.5**	15.2	**82.4**	**92.7**

Table 6
Indication of the number of parameters of every architecture used. Considering YOLOv5m6, the proposed architectures offer better results with lower parameters.

Models	Parameters(M)
YOLOv5m6 (baseline)	41.2
FRCNN	41.2
RetinaNet	34.1
FCOS	32.3
YOLO-SPAM	29.8 (−11.4)
YOLO-SPAM++	23.6 (−17.6)

subset of MP-IDB and tested on the IML dataset, YOLOv5m6 performs best in all the reported metrics.

Although the three models are in line with AP, YOLOv5m6 is the best in terms of AP50 and APL. The main motivations in this sense are related to the fact that the images of the two datasets are broadly different, and also, YOLO-SPAM models mainly aim to target small and medium-sized objects for the earliest possible detection. At the same time, IML comprises almost large parasites (see Table 2).

Moving on to the results obtained on the Pv subset of MP-IDB with models trained on IML, YOLO-SPAM++ demonstrates superior performance across all metrics, except for APL (−3.2% if compared to YOLOv5m6), confirming its promising performance on small and medium-sized objects even in a cross-dataset scenario. YOLO-SPAM struggles more than YOLO-SPAM++ in the detection of large objects, but it improves the results of YOLOv5m6 in terms of AP, AP50, and APm.

According to the findings, YOLO-SPAM++ shows potential as an effective architecture even in cross-data scenarios. However, it is important to note that YOLO-SPAM++ was explicitly designed to detect tiny parasites in their early stages. Therefore, it must be optimized for scenarios where medium or large parasites are more prevalent and established.

5.6. Qualitative analysis

Fig. 7 shows the predicted bounding boxes that our two proposed architectures have predicted. These predictions are highly accurate when compared to the ground truth. To better understand the results, we have also included the results obtained from Faster R-CNN, RetinaNet, and FCOS for comparison. During the testing phase, we found that Pf was the most difficult species to detect accurately. This difficulty was primarily due to the presence of tiny rings that represent an early infection and can be found in large numbers, with some images containing more than ten parasites. It is essential to note that this condition is unique to the Pf split. When examining Pf, it is important to consider the detection results of various detectors. YOLO-SPAM and YOLO-SPAM++ could accurately match the ground truth, while other detectors generated excess predictions. This surplus of predictions led to incorrect interpretations of certain areas, such as white blood cell nuclei being classified as parasites. This error is particularly exemplified by the FRCNN, RetinaNet, and FCOS detection results on the P. falciparum split, shown in Fig. 7(a).

YOLO-SPAM has demonstrated remarkable progress compared to our previous model, as evidenced by the findings in Fig. 6. Upon scrutinizing Fig. 6(a), it becomes apparent that our former model had difficulties detecting tiny parasites, such as early rings. However, the new architecture, notably YOLO-SPAM++, has successfully tackled this challenge. In fact, upon analyzing Fig. 6(b), every tiny ring was accurately identified, signifying a significant improvement in the task at hand.

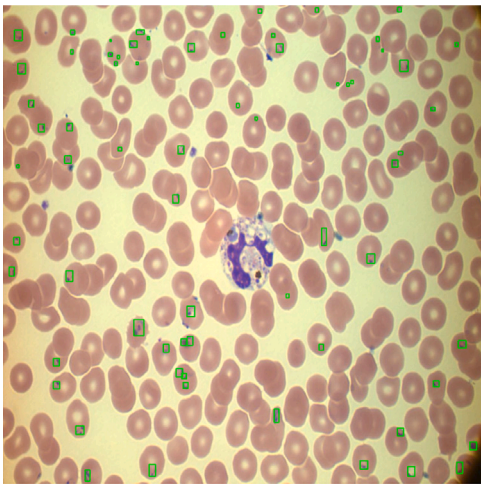
6. Discussion

The proposed method is based on the YOLOv5 object detector, with the addition of the proposed attention mechanism to focus on tiny parasites without losing details on the fully developed ones.

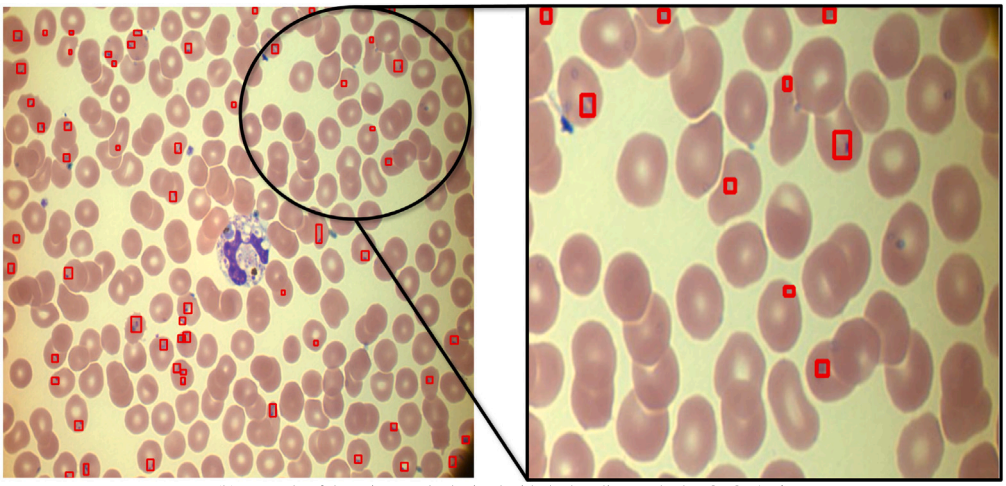
The experimental results have demonstrated that YOLO-SPAM and YOLO-SPAM++ are highly effective in parasite detection in different MP-IDB configurations and IML datasets.

As shown in Table 4, the preliminary study on the Pf subset of MP-IDB revealed that the architectural modifications significantly improved the results compared to the baseline. YOLO-SPAM++ showed a 10% improvement compared to YOLO-SPAM and up to 20% improvement against the best-performing baseline (YOLOv5l6).

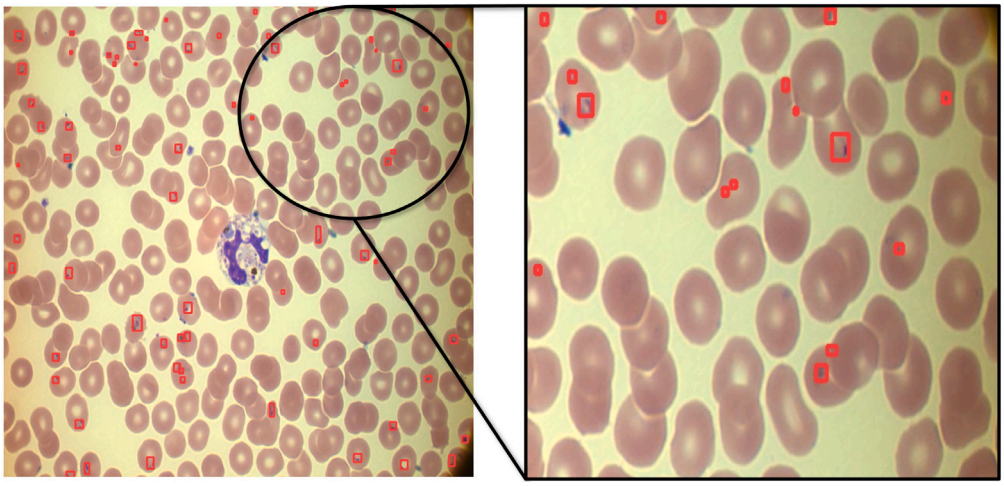
An improvement was confirmed in the next task, reported in Table 5, which involved the detection of parasites in all subsets comprising MP-IDB (Pf, Pm, Po, Pv). However, there was one relevant distinction: YOLO-SPAM produced slightly higher results than YOLO-SPAM++ in the Pm and Po subsets (+0.5% and +6% in terms of AP, respectively), which have no tiny parasites. In contrast, YOLO-SPAM++



(a) Ground truth.



(b) Example of detection result obtained with the baseline method, YOLOv5m6.



(c) Example of detection result obtained with the proposed solution, YOLO-SPAM++.

Fig. 6. Details on the improvements provided by the proposed YOLO-SPAM++ architecture. The upper section shows a sample image taken from the *P.f.* split of MP-IDB with its own ground-truth. The middle section shows the detection outcomes obtained with the baseline method, YOLOv5m6. Here, a closer look (at the right) reveals missing parasites in the detection results. In contrast, the lower section presents the results obtained with the proposed method, YOLO-SPAM++. Here, all the parasites are accurately detected, even the tiniest. This comparison underscores the enhanced precision and accuracy achieved by YOLO-SPAM++.

produced higher results in subsets containing tiny parasites like Pf and Pv. Therefore, YOLO-SPAM is proposed to be a more accurate detector of already fully developed parasites, whereas YOLO-SPAM++ sacrifices this aspect to some extent by providing results that are more focused on detecting smaller parasites, a known first sign of early infection.

We conducted cross-species generalization experiments to address the challenges posed by fully developed parasites in the context of YOLO-SPAM++ and demonstrate the robustness of the proposed architectures. This setting involved merging the training sets of various species within MP-IDB (refer to Table 7). Compared to the results

Table 7
Detection results on the four different MP-IDB species obtained with the proposed architectural variants trained with the merged training sets of the four species.

Dataset	Species	Method	AP (%)	AP_{50}	AP_s	AP_m	AP_L
MP-IDB	P. falciparum	YOLOv5m6	62.9	92.3	64.5	67.1	**100.0**
		YOLO-SPAM	68.1	95.0	70.2	71.0	**100.0**
		YOLO-SPAM++	**73.7**	**96.5**	**74.1**	**76.9**	**100.0**
	P. malariae	YOLOv5m6	82.1	94.5	–	81.7	–
		YOLO-SPAM	81.4	96.4	–	82.4	–
		YOLO-SPAM++	**84.3**	**96.5**	–	**84.0**	
	P. ovale	YOLOv5m6	81.6	90.9	–	81.6	–
		YOLO-SPAM	83.1	**94.9**	–	83.1	–
		YOLO-SPAM++	**87.9**	**94.9**	–	**87.9**	–
	P. vivax	YOLOv5m6	68.9	86.7	10.0	73.8	89.5
		YOLO-SPAM	73.1	87.7	**20.0**	76.2	90.9
		YOLO-SPAM++	**76.9**	**89.3**	11.8	**81.3**	**94.4**

Table 8
Detection results on IML dataset.

Dataset	Species	Method	AP (%)	AP_{50}	AP_s	AP_m	AP_L
IML	P. vivax	YOLOv5m6	60.4	86.5	0.0	61.7	63.0
		YOLO-SPAM	**62.0**	**89.1**	0.0	**64.2**	64.2
		YOLO-SPAM++	61.5	86.9	0.0	57.7	**64.4**

Table 9
Experimental results obtained with the cross-dataset scenario. The first column represents the results obtained with the models trained on the training set of the Pv subset of MP-IDB and tested on the test set of IML. The second column shows the results with the same models trained on the training set of IML and tested on the test set of the Pv split of MP-IDB.

Method	Train: MP-IDB (P. vivax)					Train: IML				
	Test: IML					Test: MP-IDB (P. vivax)				
	AP (%)	AP_{50}	AP_s	AP_m	AP_L	AP (%)	AP_{50}	AP_s	AP_m	AP_L
YOLOv5m6	**13.7**	**29.1**	0.0	**3.1**	**21.1**	7.2	16.9	0.0	6.7	24.0
YOLO-SPAM	11.4	24.3	0.0	2.6	15.4	7.4	17.0	0.0	6.9	12.4
YOLO-SPAM++	10.1	18.6	0.0	1.0	14.0	**8.6**	**19.4**	0.0	**9.5**	**20.8**

obtained from the previous configuration, YOLO-SPAM++ consistently outperformed the YOLOv5 baseline and YOLO-SPAM across all subsets. These results highlight how including diverse representations of different parasite species contributed to the model achieving superior performance while maintaining high precision on tiny parasites. Notably, the relatively low detection rate of tiny parasites in the Pv subset could be attributed to their limited representation, introducing additional complexities to the detection process.

To further validate the proposed architectures, they were tested on an external dataset, i.e., IML, and in a cross-dataset scenario, as indicated in Table 8 and Table 9, respectively. Again, the improvement provided by YOLO-SPAM in both its architectures over the baseline was consistent. In the first case, YOLO-SPAM outperformed the baseline provided by YOLOv5 and YOLO-SPAM++ by 0.5% and 6.5% in terms of AP and AP_m, respectively. This result is motivated by the absence of tiny parasites in the composition of IML, leading YOLO-SPAM++ to some issues when generalization is needed.

When testing cross-datasets, YOLO-SPAM++ performs better when trained on IML and tested on the Pv partition of MP-IDB. However, when the Pv partition of MP-IDB is used as the training set and IML as the testing set, YOLO-SPAM and YOLO-SPAM++ show lower results than the baseline provided by YOLOv5. This result is understandable, given the wide range of variations present in MP-IDB from multiple perspectives, such as staining, illumination, and Pv life stage composition, compared to what is present in IML. Under these circumstances, the proposed specialized architectures face more significant challenges in detecting parasites.

From a general point of view, however, the proposed architectures are suitable for detecting malaria parasites from full-size images, especially when dealing with small-size parasites.

A final significant improvement over state of the art is represented in Table 6. As can be seen, the performance improvements are to be considered even more consistent because the two proposed architectures have significantly fewer parameters than YOLOv5m6, i.e., the architecture considered as the baseline. Specifically, YOLO-SPAM has 11.4M fewer parameters than the baseline, while YOLO-SPAM++ has 17.6M.

Limitations. Although the two types of architecture improved the task, they exhibit some limitations. Firstly, detecting small parasites was successful on the Pf subset, but the results on the P. vivax subset were unsatisfactory due to their low representation. To address this, techniques like class imbalance or few-shot learning could be used in the future. Secondly, there was no clear best in some cases, as different approaches performed better depending on the subset. For example, YOLO-SPAM++ worked best on Pf and Pv subsets, while YOLO-SPAM was better on Pm and Po. This variation could be because, having only two prediction heads, YOLO-SPAM++ struggles with limited samples as in Pm and Po.

Additionally, Fig. 8 graphically represents how the proposed architectures exhibit some issues, exemplified by several illustrations. Fig. 8(a) shows that only one parasite is correctly detected. On the other hand, Fig. 8(e) displays an additional parasite detected in the top border. In this case, we pinpoint that a postprocessing step may be implemented to clean the border. However, for fairness, we left the detectors as implemented. In addition, YOLO-SPAM accurately pinpoints the parasite in Fig. 8(b), but YOLO-SPAM++ fails to detect the same parasite in Fig. 8(f). Conversely, in Figs. 8(c) and 8(g), YOLO-SPAM++ identifies all of them, while YOLO-SPAM misses the correct detection. Lastly, Figs. 8(d) and 8(h) demonstrate that both models struggle with some WBC nuclei components by detecting them as parasites.

Finally, moving to the cross-dataset results, three main limitations were found. First, the images of MP-IDB and IML are largely different between them. This aspect causes a first domain shift. Second, YOLO-SPAM-based models mainly aim to target small and medium-sized

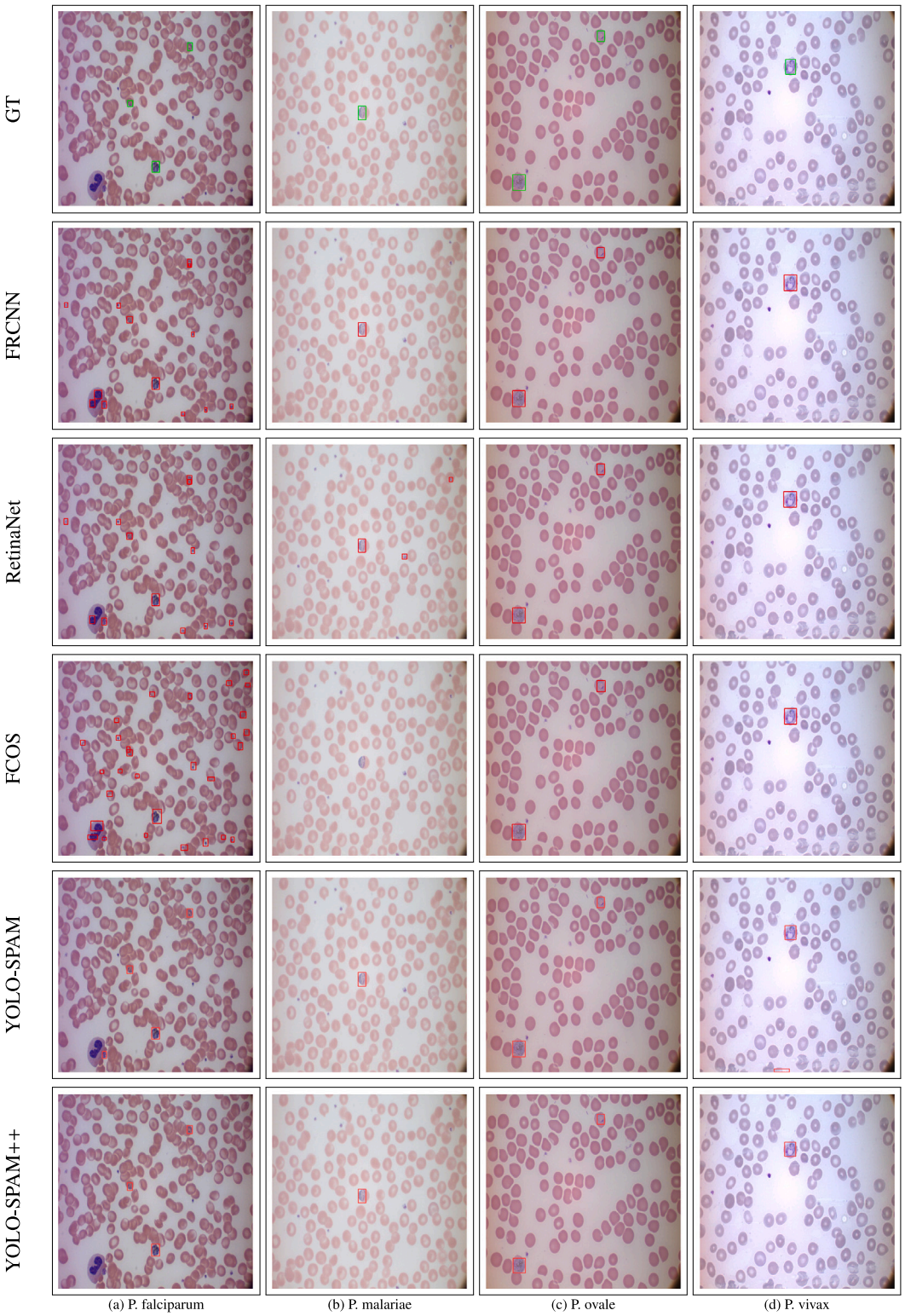


Fig. 7. Each column represents a different malaria species included in MP-IDB. From left to right: Pf, Pm, Po, Pv. The first row (GT) represents the ground truth. Every subsequent row represents the results obtained by the detection methods used for comparison purposes. From the second to the last row: FRCNN, RetinaNet, FCOS, YOLO-SPAM, and YOLO-SPAM++. Finally, □ represent the ground truth, while □ indicates the detected parasites.

objects for the earliest possible detection, but IML is composed almost of large objects, as indicated in Table 2. Therefore, this aspect causes the second structural domain shift. Third, certain critical aspects were found to be problematic for detection in the IML dataset. These issues are highlighted in Fig. 9, where object detectors can be misled by bounding boxes that encompass areas beyond the parasites.

7. Conclusions

The two architectures of the malaria parasite detector proposed in this study make a substantial contribution to the detection of malaria.

Recalling that the main objective of this work was to address the problem of detecting tiny parasites for early diagnosis, we implemented

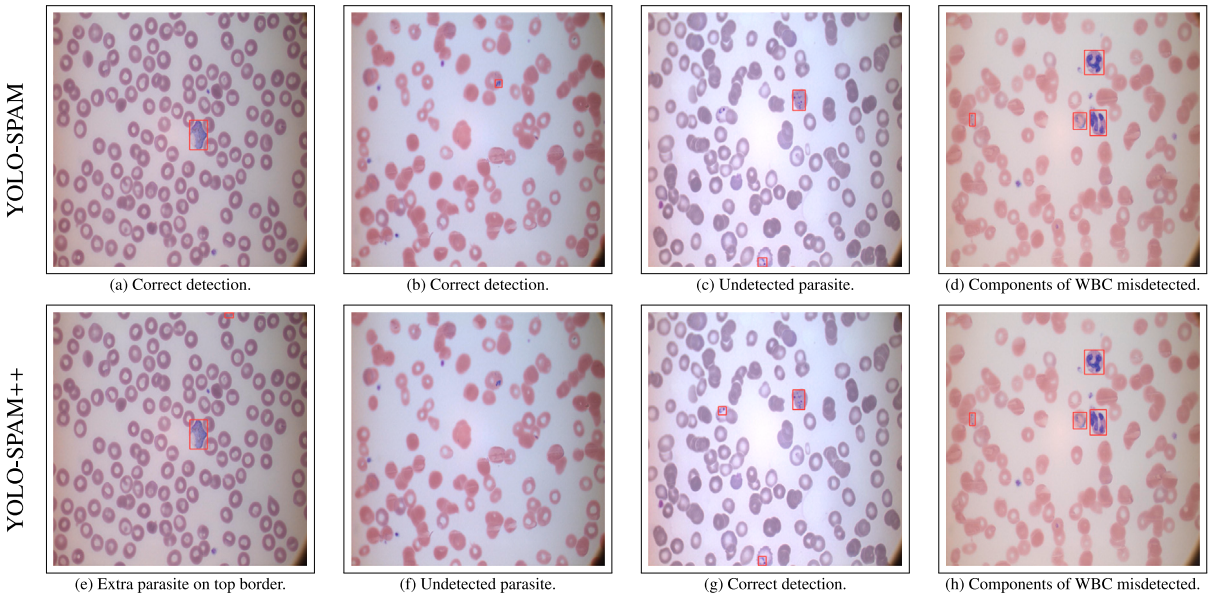


Fig. 8. Examples of detection issues found in the proposed architectures. □ indicates the detected parasites. Fig. 8(a) shows the correct detection of the only parasite present, while Fig. 8(e) shows a further parasite detected in the top border. Fig. 8(b) shows the correct detection by YOLO-SPAM. In contrast, Fig. 8(f) represents the same parasite, undetected by YOLO-SPAM++. Fig. 8(c) and Fig. 8(g) show the same problem in reverse parts, with YOLO-SPAM missing the correct detection and YOLO-SPAM++ identifying all of them. Finally, Figs. 8(d) and 8(h) show that, in some cases, both models detect some components of WBC nuclei as parasites.

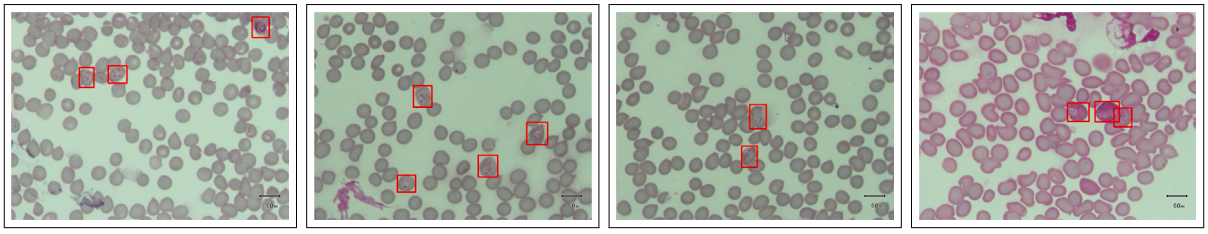


Fig. 9. Examples of critical aspects found in the ground truth provided with IML. All the images show bounding box including parts of further regions that can challenge the detectors with extra zones and edges.

two new architectures that offered promising results and improved state of the art in terms of AP and APs. Both architectures possess fewer parameters than the baseline considered, adapting to the use case represented by low-end devices.

Benchmarking two public datasets has demonstrated that the proposed approach is highly effective and superior to existing state-of-the-art methods. This result is achieved through multiple attention mechanisms that solve the problem of detecting tiny parasites, a significant challenge current methods face.

The study unveils that the proposed architectures demonstrate outstanding performance in identifying malaria parasites in diverse situations, including the detection of multiple species simultaneously. Furthermore, the outcomes of cross-dataset experiments are also encouraging despite the challenges faced during the procedure.

Moving forward, we have identified various research initiatives we intend to pursue. Our overarching objective is to enhance our methodology to enable us to detect all types of malaria parasites with greater precision. While our current dataset has yielded encouraging outcomes, we are eager to refine our system to operate on a cross-dataset model using, for instance, synthetic images produced with generative adversarial networks or diffusion models. Doing so will enable it to effectively contend with environmental variances between varying datasets. Furthermore, our ultimate aspiration is to extend our approach to encompass a multi-magnification image representation of the same blood smear. This aspect will allow us to more accurately identify malaria parasites across differing magnifications.

CRediT authorship contribution statement

Luca Zedda: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Writing – original draft, Writing – review & editing. **Andrea Loddo:** Conceptualization, Data curation, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Cecilia Di Ruberto:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset used are publicly available and referenced in the manuscript.

Acknowledgment

This work was supported in part by the Fondazione di Sardegna, Italy Project F72F20000190007 "Analysis of innovative Blockchain technologies: Libra, Bitcoin and Ethereum and technological, economical and social comparison among these different blockchain technologies".

References

[1] World Health Organization, 2023, https://www.who.int/news-room/fact-sheets/detail/malaria. (Online; Accessed 29 May 2023).
[2] Stanford Healthcare, 2021, https://stanfordhealthcare.org/medical-conditions/primary-care/malaria/types.html. (Online; Accessed 29 May 2023).
[3] United States' Centers for Disease Control and Prevention, 2021, https://www.cdc.gov/malaria/about/biology/index.html. (Online; Accessed 29 May 2023).
[4] A.M. Gimenez, R.F. Marques, M. Regiart, D.Y. Bargieri, Diagnostic methods for non-falciparum malaria, Front. Cell. Infect. Microbiol. 11 (2021) 681063.
[5] A. Vijayalakshmi, B.R. Kanna, Deep learning approach to detect malaria from microscopic images, Multim. Tools Appl. 79 (21–22) (2020) 15297–15317.
[6] A. Loddo, C. Di Ruberto, M. Kocher, G. Prod'Hom, MP-IDB: the malaria parasite image database for image processing and analysis, in: N. Leporé, J. Brieva, E. Romero, D. Racoceanu, L. Joskowicz (Eds.), Processing and Analysis of Biomedical Information - First International SIPAIM Workshop, SaMBa 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Revised Selected Papers, in: Lecture Notes in Computer Science, vol. 11379, Springer, 2018, pp. 57–65.
[7] M. Maity, A. Jaiswal, K. Gantait, J. Chatterjee, A. Mukherjee, Quantification of malaria parasitaemia using trainable semantic segmentation and capsnet, Pattern Recognit. Lett. 138 (2020) 88–94.
[8] P. Berzosa, A. de Lucio, M. Romay-Barja, Z. Herrador, V. González, L. García, A. Fernández-Martínez, M. Santana-Morales, P. Ncogo, B. Valladares, et al., Comparison of three diagnostic methods (microscopy, RDT, and PCR) for the detection of malaria parasites in representative samples from Equatorial Guinea, Malaria J. 17 (1) (2018) 1–12.
[9] A. Onken, C.G. Haanshuus, M.K. Miraji, M. Marijani, K.O. Kibwana, K.A. Abeid, K. Mørch, M. Reimers, N. Langeland, F. Müller, et al., Malaria prevalence and performance of diagnostic tests among patients hospitalized with acute undifferentiated fever in Zanzibar, Malar. J. 21 (1) (2022) 1–8.
[10] Q. Huang, W. Li, B. Zhang, Q. Li, R. Tao, N.H. Lovell, Blood cell classification based on hyperspectral imaging with modulated gabor and CNN, IEEE J. Biomed. Health Inf. 24 (1) (2020) 160–170.
[11] L.H.S. Vogado, R.M.S. Veras, F.H.D. Araújo, R.R.V. e Silva, K.R.T. Aires, Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification, Eng. Appl. Artif. Intell. 72 (2018) 415–422.
[12] M. Togaçar, B. Ergen, Z. Cömert, Classification of white blood cells using deep features obtained from Convolutional Neural Network models based on the combination of feature selection methods, Appl. Soft Comput. 97 (Part B) (2020) 106810.
[13] C. Di Ruberto, A. Loddo, L. Putzu, Detection of red and white blood cells from microscopic blood images using a region proposal approach, Comput. Biol. Med. 116 (2020) 103530.
[14] W. Xie, J.A. Noble, A. Zisserman, Microscopy cell counting and detection with fully convolutional regression networks, Comput. methods Biomech. Biomed. Eng. Imaging Vis. 6 (3) (2018) 283–292, http://dx.doi.org/10.1080/21681163.2016.1149104.
[15] Z. Liang, A. Powell, I. Ersoy, M. Poostchi, K. Silamut, K. Palaniappan, P. Guo, M.A. Hossain, S.K. Antani, R.J. Maude, J.X. Huang, S. Jaeger, G.R. Thoma, CNN-based image analysis for malaria diagnosis, in: IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016, Shenzhen, China, December 15-18, 2016, IEEE Computer Society, 2016, pp. 493–496.
[16] S. Rajaraman, S.K. Antani, M. Poostchi, K. Silamut, M.A. Hossain, R.J. Maude, S. Jaeger, G.R. Thoma, Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images, PeerJ 6 (2018) e4568.
[17] S. Rajaraman, S. Jaeger, S.K. Antani, Perf. eval. of deep neural ensembles toward malaria parasite detection in thin-blood smear images, PeerJ 7 (2019) e6977.
[18] C. Di Ruberto, A. Dempster, S. Khan, B. Jarra, Analysis of infected blood cell images using morphological operators, Image Vision Comput. 20 (2) (2002) 133–146.
[19] F.B. Tek, A.G. Dempster, I. Kale, Malaria parasite detection in peripheral blood images, in: M.J. Chantler, R.B. Fisher, E. Trucco (Eds.), Proceedings of the British Machine Vision Conference 2006, Edinburgh, UK, September 4-7, 2006, British Machine Vision Association, 2006, pp. 347–356.
[20] S.K. Kumarasamy, S. Ong, K.S. Tan, Robust contour reconstruction of red blood cells and parasites in the automated identification of the stages of malarial infection, Mach. Vis. Appl. 22 (3) (2011) 461–469.
[21] S. Bias, S. Reni, I. Kale, Mobile hardware based implementation of a novel, efficient, fuzzy logic inspired edge detection technique for analysis of malaria infected microscopic thin blood images, in: The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018) / the 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops, November 5-8, 2018, Leuven, Belgium, in: Procedia Computer Science, vol. 141, Elsevier, 2018, pp. 374–381.
[22] A. Loddo, L. Putzu, On the effectiveness of leukocytes classification methods in a real application scenario, AI 2 (3) (2021) 394–412.
[23] W. Hu, C. Li, X. Li, M.M. Rahaman, J. Ma, Y. Zhang, H. Chen, W. Liu, C. Sun, Y. Yao, et al., GasHisSDB: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer, Comput. Biol. Med. 142 (2022) 105207.
[24] J. Gayathri, B. Abraham, M. Sujarani, M.S. Nair, A computer-aided diagnosis system for the classification of COVID-19 and non-COVID-19 pneumonia on chest X-ray images by integrating CNN with sparse autoencoder and feed forward neural network, Comput. Biol. Med. 141 (2022) 105134.
[25] H. Li, N. Zeng, P. Wu, K. Clawson, Cov-Net: A computer-aided diagnosis method for recognizing COVID-19 from chest X-ray images via machine vision, Expert Syst. Appl. 207 (2022) 118029.
[26] M. Zaid, S. Ali, M. Ali, S. Hussein, A. Saadia, W. Sultani, Identifying out of distribution samples for skin cancer and malaria images, Biomed. Signal Process. Control 78 (2022) 103882.
[27] W. Sultani, W. Nawaz, S. Javed, M.S. Danish, A. Saadia, M. Ali, Towards low-cost and efficient malaria detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 20655–20664.
[28] L. Zedda, A. Loddo, C. Di Ruberto, A deep learning based framework for malaria diagnosis on high variation data set, in: Image Analysis and Processing - ICIAP 2022 - 21st International Conference, Lecce, Italy, May 23-27, 2022, Proceedings, Part II, in: Lecture Notes in Computer Science, vol. 13232, Springer, 2022, pp. 358–370.
[29] A. Loddo, C. Di Ruberto, M. Kocher, Recent advances of malaria parasites detection systems based on mathematical morphology, Sensors 18 (2) (2018) 513.
[30] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. of the 25th International Conference on Neural Information Processing Systems, Vol. 1, NIPS '12, 2012, pp. 1097–1105.
[31] Q.A. Arshad, M. Ali, S. Hassan, C. Chen, A. Imran, G. Rasul, W. Sultani, A dataset and benchmark for malaria life-cycle classification in thin blood smear images, Neural Comput. Appl. 34 (6) (2022) 4473–4485.
[32] S. Marletta, V. L'Imperio, A. Eccher, P. Antonini, N. Santonicco, I. Girolami, A.P. Dei Tos, M. Sbaraglia, F. Pagni, M. Brunelli, et al., Artificial intelligence-based tools applied to pathological diagnosis of microbiological diseases, Pathol.-Res. Pract. (2023) 154362.
[33] A. Loddo, L. Putzu, On the reliability of CNNs in clinical practice: a computer-aided diagnosis system case study, Appl. Sci. 12 (7) (2022) 3269.
[34] C. Lin, H. Wu, Z. Wen, J. Qin, Automated malaria cells detection from blood smears under severe class imbalance via importance-aware balanced group softmax, in: M. de Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part VIII, in: Lecture Notes in Computer Science, vol. 12908, Springer, 2021, pp. 455–465.
[35] P. Manescu, C. Bendkowski, R. Claveau, M. Elmi, B.J. Brown, V. Pawar, M.J. Shaw, D. Fernandez-Reyes, A weakly supervised deep learning approach for detecting malaria and sickle cells in blood films, in: A.L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part V, in: Lecture Notes in Computer Science, vol. 12265, Springer, 2020, pp. 226–235.
[36] F. Abdurahman, K.A. Fante, M. Aliy, Malaria parasite detection in thick blood smear microscopic images using modified YOLOV3 and YOLOV4 models, BMC Bioinform. 22 (1) (2021) 112.
[37] S. Chibuta, A.C. Acar, Real-time malaria parasite screening in thick blood smears for low-resource setting, J. Dig. Imaging 33 (3) (2020) 763–775.
[38] A. Koirala, M. Jha, S. Bodapati, A. Mishra, G. Chetty, P.K. Sahu, S. Mohanty, T.K. Padhan, J. Mattoo, A. Hukkoo, Deep learning for real-time malaria parasite detection and counting using YOLO-mp, IEEE Access 10 (2022) 102157–102172.
[39] A. Rahman, H. Zunair, T.R. Reme, M.S. Rahman, M. Mahdy, A comparative analysis of deep learning architectures on high variation malaria parasite classification dataset, Tissue Cell 69 (2021) 101473.
[40] A. Diker, An efficient model of residual based convolutional neural network with Bayesian optimization for the classification of malarial cell images, Comput. Biol. Med. (2022) 105635.
[41] N. Sengar, R. Burget, M.K. Dutta, A vision transformer based approach for analysis of plasmodium vivax life cycle for malaria prediction using thin blood smear microscopic images, Comput. Methods Programs Biomed. 224 (2022) 106996, http://dx.doi.org/10.1016/j.cmpb.2022.106996.

[42] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, CVPR, 2009, pp. 248–255.
[43] H. Guan, M. Liu, Domain adaptation for medical image analysis: A survey, IEEE Trans. Biomed. Eng. 69 (3) (2022) 1173–1185, http://dx.doi.org/10.1109/TBME.2021.3117407.
[44] F. Yang, M. Poostchi, H. Yu, Z. Zhou, K. Silamut, J. Yu, R.J. Maude, S. Jäger, S.K. Antani, Deep learning for smartphone-based malaria parasite detection in thick blood smears, IEEE J. Biomed. Health Inform. 24 (5) (2020) 1427–1438.
[45] Q.A. Arshad, M. Ali, S. Hassan, C. Chen, A. Imran, G. Rasul, W. Sultani, A dataset and benchmark for malaria life-cycle classification in thin blood smear images, Neural Comput. Appl. 34 (6) (2022) 4473–4485.
[46] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, 2015, pp. 91–99.
[47] T. Lin, P. Goyal, R.B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2999–3007.
[48] Z. Tian, C. Shen, H. Chen, T. He, FCOS: fully convolutional one-stage object detection, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 9626–9635.
[49] J. Redmon, S.K. Divvala, R.B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 779–788.
[50] H. Zhou, F. Jiang, H. Lu, SSDA-YOLO: semi-supervised domain adaptive YOLO for cross-domain object detection, 2022, CoRR abs/2211.02213,
[51] Z. Zou, K. Chen, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A survey, Proc. IEEE 111 (3) (2023) 257–276.
[52] T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, in: Lecture Notes in Computer Science, vol. 8693, Springer, 2014, pp. 740–755.
[53] X. Zhu, S. Lyu, X. Wang, Q. Zhao, TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios, in: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021, IEEE, 2021, pp. 2778–2788.
[54] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[55] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 8759–8768.
[56] S. Woo, J. Park, J. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, in: Lecture Notes in Computer Science, vol. 11211, Springer, 2018, pp. 3–19.
[57] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.
[58] M. Guo, T. Xu, J. Liu, Z. Liu, P. Jiang, T. Mu, S. Zhang, R.R. Martin, M. Cheng, S. Hu, Attention mechanisms in computer vision: A survey, Comput. Vis. Media 8 (3) (2022) 331–368.
[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008.
[60] yichao liu, Z. Shao, yueyang Teng, N. Hoffmann, NAM: Normalization-based attention module, in: NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future, 2021, URL https://openreview.net/forum?id=AaTK_ESdkjg.
[61] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 9992–10002.
[62] H. Gong, T. Mu, Q. Li, H. Dai, C. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, H. Li, X. Lang, Z. Li, B. Wang, Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images, Remote Sens. 14 (12) (2022) 2861.
[63] G. Jocher, et al., ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, Zenodo, 2021, http://dx.doi.org/10.5281/zenodo.5563715.
[64] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, 2019, https://github.com/facebookresearch/detectron2.
[65] R. Padilla, S.L. Netto, E.A.B. da Silva, A survey on performance metrics for object-detection algorithms, in: 2020 International Conference on Systems, Signals and Image Processing, IWSSIP 2020, NiterÓI, Brazil, July 1-3, 2020, IEEE, 2020, pp. 237–242.