MalariaSD: Malaria-Infected Cell Images dataset

Shafina Charania ¹ and Ninad Mehendale ²

¹Affiliation not available
²K J Somaiya College of Engineering

October 31, 2023

Abstract

MalariaSD is a dataset encompassing various stages and classes of the malaria parasite, including Plasmodium falciparum, Plasmodium malariae, Plasmodium vivax, and Plasmodium ovale. These divisions consist of four distinct phases: ring, schizont, trophozoite, and gametocyte stages. The dataset serves as a valuable resource for researchers and healthcare professionals, offering crucial insights into the epidemiology, diagnosis, and treatment of malaria. The MP-IDB, a comprehensive collection of high-quality malaria parasite images, features the aforementioned four stages. This database presents an opportunity to develop and evaluate novel image processing and analysis techniques, aiming to enhance the accuracy and efficiency of malaria diagnosis. In our proposed paper, these images were used to create a new dataset using stable diffusion and advanced image processing methods. By utilizing stable diffusion, we generated a dataset comprising 16 distinct classes. Specifically, we focused on single-celled images and applied cropping and enhancement techniques to produce refined images. Subsequently, this new dataset underwent training through stable diffusion, resulting in the generation of 20 additional images for each class. As a result of our efforts, the image count of the original dataset increased significantly from an average of 12 images to 40 images per class. Through the expansion of the dataset using stable diffusion and image processing, our paper contributes to the advancement of malaria research. The augmented dataset provides a more comprehensive representation of the various stages and classes of malaria parasites, empowering researchers and healthcare professionals to enhance their understanding of malaria’s complexities and improve diagnostic methodologies.

MalariaSD is a dataset encompassing various stages and classes of the malaria parasite, including Plasmodium falciparum, Plasmodium malariae, Plasmodium vivax, and Plasmodium ovale. These divisions consist of four distinct phases: ring, schizont, trophozoite, and gametocyte stages. The dataset serves as a valuable resource for researchers and healthcare professionals, offering crucial insights into the epidemiology, diagnosis, and treatment of malaria.

The MP-IDB, a comprehensive collection of high-quality malaria parasite images, features the aforementioned four stages. This database presents an opportunity to develop and evaluate novel image processing and analysis techniques, aiming to enhance the accuracy and efficiency of malaria diagnosis. In our proposed paper, these images were used to create a new dataset using stable diffusion and advanced image processing methods.

By utilizing stable diffusion, we generated a dataset comprising 16 distinct classes. Specifically, we focused on single-celled images and applied cropping and enhancement techniques to produce refined images. Subsequently, this new dataset underwent training through stable diffusion, resulting in the generation of 20 additional images for each class. As a result of our efforts, the image count of the original dataset increased significantly from an average of 12 images to 40 images per class.

Through the expansion of the dataset using stable diffusion and image processing, our paper contributes to the advancement of malaria research. The augmented dataset provides a more comprehensive representation of the various stages and classes of malaria parasites, empowering researchers and healthcare professionals to enhance their understanding of malaria’s complexities and improve diagnostic methodologies.
MalariaSD: Malaria-Infected Cell Images dataset

Shafina Charania and Ninad Mehendale

Index Terms—Computer Vision, image Processing, Stable Diffusion, Malaria Cells

I. ABSTRACT

MalariaSD is a dataset encompassing various stages and classes of the malaria parasite, including Plasmodium falciparum, Plasmodium malariae, Plasmodium vivax, and Plasmodium ovale. These divisions consist of four distinct phases: ring, schizont, trophozoite, and gametocyte stages. The dataset serves as a valuable resource for researchers and healthcare professionals, offering crucial insights into the epidemiology, diagnosis, and treatment of malaria.

The MP-IDB [1], a comprehensive collection of high-quality malaria parasite images, features the aforementioned four stages. This database presents an opportunity to develop and evaluate novel image processing and analysis techniques, aiming to enhance the accuracy and efficiency of malaria diagnosis. In our proposed paper, these images were used to create a new dataset using stable diffusion and advanced image processing methods.

By utilizing stable diffusion, we generated a dataset comprising 16 distinct classes. Specifically, we focused on single-celled images and applied cropping and enhancement techniques to produce refined images. Subsequently, this new dataset underwent training through stable diffusion, resulting in the generation of 20 additional images for each class. As a result of our efforts, the image count of the original dataset increased significantly from an average of 12 images to 40 images per class.

Through the expansion of the dataset using stable diffusion and image processing, our paper contributes to the advancement of malaria research. The augmented dataset provides a more comprehensive representation of the various stages and classes of malaria parasites, empowering researchers and healthcare professionals to enhance their understanding of malaria’s complexities and improve diagnostic methodologies.

II. INTRODUCTION

For effective treatment of malaria, a potentially fatal disease spread by Plasmodium parasites, prompt detection and precise classification are essential. However, the availability of a well-balanced dataset with high-resolution images encompassing all malaria classes is essential for robust classification models. The existing "MP-IDB" dataset, although valuable, suffers from multicolored, multi-class images and an imbalance in class representation.

To address these limitations and advance malaria classification research, we have developed a novel dataset. By incorporating images from the "MP-IDB" dataset and additional images from Google Search, we aimed to achieve class balance. Through rigorous preprocessing, including noise elimination and maintaining image quality, we created a uniform dataset.

The dataset exhibits potential for training machine learning models to achieve precise classification of malaria parasites within blood smear images. Through addressing class imbalances and enhancing image quality, the objective is to enhance the dependability and effectiveness of malaria diagnosis algorithms.

In this research paper, we present the dataset creation process, including data collection methodology, preprocessing techniques, and class balance. Our research contributes to malaria diagnosis by providing a comprehensive and balanced dataset, enabling more accurate and efficient parasite classification. By overcoming the limitations of existing datasets, we aim to enhance automated systems’ capabilities in detecting and identifying malaria infections, ultimately improving patient care and management.

A. Malaria and its different stages

The transformation of the malaria parasite from one stage to another involves a complex life cycle within both the human host and the mosquito vector. The stages of the malaria parasite’s life cycle are as follows:

1. Ring Stage: The malaria parasite enters the human host through the bite of an infected mosquito. The parasite initially appears in the form of a ring stage, which has a small size and a single nucleus. It invades red blood cells and starts to multiply within them.

2. Trophozoite Stage: As the parasite matures, it progresses to the trophozoite stage. During this stage, the parasite grows in size and develops multiple nuclei. It continues to consume nutrients from the host red blood cells, contributing to the destruction of these cells.

3. Schizont Stage: The trophozoite stage transitions into the schizont stage. At this point, the parasite undergoes a process known as schizogony or asexual replication. The parasite replicates its genetic material and divides into numerous daughter cells called merozoites. The host red blood cell is ruptured, releasing these merozoites into the bloodstream, where they can invade new red blood cells and continue the cycle.

4. Gametocyte Stage: Some of the parasites in the schizont stage differentiate into sexual forms known as gametocytes. These gametocytes are not involved in asexual replication but serve as the sexual stages of the parasite’s life cycle. If a mosquito takes a blood meal from an infected human, it ingests the gametocytes along with the blood.

Shafina Charania, Ninad Mehendale are with K J Somaiya College of Engineering, e-mail: shafina.c@somaiya.edu, e-mail: ninad@somaiya.edu

Ninad Mehendale is the corresponding author.
Once inside the mosquito, the gametocytes develop further, maturing into male and female gametes. Fertilization occurs, resulting in the formation of a zygote. The zygote undergoes a series of transformations, ultimately developing into sporozoites, which migrate to the mosquito’s salivary glands. From there, the mosquito can transmit the sporozoites to another human during a subsequent blood meal, restarting the cycle.

Overall, the transformation of the malaria parasite from one stage to another involves intricate developmental changes within both the human host and the mosquito vector, allowing the parasite to adapt and propagate itself.

III. METHODOLOGY

**ovale_schizont**

Our dataset consists of two types of images:

- Cropped, resized and AI-enhanced images from the IMDP dataset Stable Diffusion generated images and their respective ckpt files.

To create the first type of images, we manually did the cropping according to the label mentioned in the original dataset and then segregated them into 16 different folders according to their classtype_stage.

As a result, 16 unbalanced folders were created, and to balance them, we used images available directly on Google search or those used by other people in their works).

Thus, by doing this, we were able to create an almost balanced dataset containing about 7–20 images in each class.

We used Stable Diffusion instead of GAN models because, Stable diffusion is a type of generative model that can be trained on small datasets. It is more stable and easier to train than GANs, making it a better choice for beginners. Stable diffusion is also more efficient, as it can generate images much faster than GANs. This makes it a better choice for real-time applications. Additionally, stable diffusion is more versatile than GANs, as it can be used to generate a wider variety of images. Overall, stable diffusion is a better choice than GAN for small datasets.

The MP-IDB (Malaria Parasite Image Database) is a public image dataset of thin blood smears infected by the malaria parasite. It comprises four species of malaria parasites: Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale, and Plasmodium malariae. For each species, there are four distinct stages of life: ring, trophozoite, schizont, and gametocyte.

The dataset contains a total of 210 images, with the following distribution:

- Plasmodium falciparum: 70 images (33 ring, 24 trophozoite, 13 schizont, and 0 gametocyte)
- Plasmodium vivax: 40 images (20 ring, 10 trophozoite, 10 schizont, and 0 gametocyte)
- Plasmodium ovale: 40 images (20 ring, 10 trophozoite, 10 schizont, and 0 gametocyte)
- Plasmodium malariae: 60 images (30 ring, 20 trophozoite, 10 schizont, and 0 gametocyte)

The images in the MP-IDB were acquired using a Leica DM2000 optical laboratory microscope with 100x magnification. They are all in JPEG format and have a resolution of 1280 x 1024 pixels.

Dataset preparation steps:

- This is the modified updated dataset, which has images from the cropped original dataset and Google search.
- All these images were then cropped using a 1:1 ratio on an online tool called Bulkcrop.
- These images were then fed to an AI-powered resolution enhancer app called GIGAPIXEL.
- It changed the resolution of these images by a significant amount.
- They were also resized to 512 by 512 pixels to create uniformity.

The drawback of this app was that it was available for a free trial and resulted in printing a watermark on the image.

An online AI tool by the name of PixelBin further removed this.

But even this tool had its drawbacks, as it allowed only 30 images to be processed.

Thus, we created new accounts and completed the watermark removal process.

These 16 folders were then fed to the stable diffusion network one by one, which resulted in generating about 20 new images per class by changing the output parameters like num_samples, guidance_scale, and num_inference_steps.

Stable diffusion is a text-to-image diffusion model that enables the generation of high-quality images by leveraging text prompts. The utilization of stable diffusion involves providing a descriptive text prompt that outlines the desired image. Subsequently, the model undergoes a series of iterations, generating a sequence of images that progressively converge towards the target image. Finally, users can select the image that best aligns with their preferences from the generated sequence.

The steps to work with stable diffusion to train a custom dataset are:

- Download the Stable Diffusion code and install it. Prepare your dataset of images. The images should be in a Python-readable format, such as JPEG or PNG.
- Train the model on your dataset. This can take a few hours or even days, depending on the size of your dataset.
- Generate images from the model. You can use the model to generate images from text prompts.

These are the most commonly used parameters once you have trained the model:

- num_samples: The number of samples used in stable diffusion, representing the quantity of data points generated during the diffusion process. guidance_scale: A parameter controlling the strength of guidance provided to the diffusion process, influencing the direction and behavior of the diffusion process. num_inference_steps: The number of inference steps taken during the stable diffusion process, indicating the iterations or stages involved in achieving stable results.

Since we had about 8 (minimum) and 20 (maximum) images per class before stable diffusion, we got about 40 images (table diffusion generated and original images) out of the model, and if we had more data, the generated images would be closer to reality and also hold ground.
IV. IMPLEMENTATION

Following are the images of the final datasets after cropping, enhancing and applying Standard Diffusion:

V. FUTURE SCOPE

The development and utilization of malaria multiclass and multistage image datasets have immense potential for various applications in malaria detection, diagnosis, and treatment. Looking ahead, several avenues of future research and exploration can be pursued to further advance the field:

1. Refinement of Machine Learning Algorithms: Continued refinement and optimization of machine learning and deep learning algorithms can improve the accuracy and efficiency of automated malaria diagnosis. Future research can focus on developing advanced models that can handle larger datasets, accommodate different imaging modalities, and effectively classify diverse types and stages of malaria parasites.

2. Integration of Clinical Data: Integrating clinical data, such as patient demographics, symptoms, and treatment history, with the malaria image dataset can provide a comprehensive understanding of the disease. This integration can enable the development of holistic diagnostic models and treatment recommendations based on a combination of clinical and imaging data.
3. Real-time Monitoring and Point-of-Care Applications: Future research can explore the development of real-time monitoring systems and point-of-care applications using the malaria image dataset. Portable devices equipped with image analysis capabilities can assist healthcare providers in making rapid and accurate malaria diagnoses at the point of care, particularly in resource-limited settings.

4. Longitudinal Studies and Treatment Response Analysis: Conducting longitudinal studies using the malaria image dataset can shed light on the progression of the disease and the efficacy of different treatment regimens. By analyzing the
changes in parasite types and stages over time, researchers can gain insights into treatment response patterns, drug resistance, and the development of targeted therapies.

5. Data Sharing and Collaboration: Promoting data sharing and collaboration among researchers and institutions can accelerate advancements in malaria research. Encouraging the open sharing of annotated malaria image datasets can foster the development of standardized evaluation benchmarks, comparative studies, and collaborative efforts aimed at tackling the challenges of malaria detection and treatment.

6. AI-driven Decision Support Systems: Expanding the capabilities of AI-driven decision support systems in malaria management can revolutionize healthcare practices. Future research can explore the integration of the malaria image dataset
with patient electronic health records, genomic data, and environmental factors to develop comprehensive AI models that provide personalized treatment recommendations, predict disease outbreaks, and optimize resource allocation.

By pursuing these future research directions, we can harness the full potential of malaria multiclass and multistage image datasets, paving the way for improved malaria diagnosis, treatment, and prevention strategies and ultimately making significant strides in combating this global health burden.

VI. CONCLUSION:

In conclusion, the development of the malaria dataset, which encompasses multiple stages and classes of the malaria parasite, offers significant opportunities for advancements in malaria research. Through stable diffusion and advanced image processing methods, we have expanded the dataset, achieving a more balanced representation of parasite classes. This augmented dataset holds promise for various applications, including the development of accurate machine-learning algorithms for automated diagnosis, drug evaluation, treatment monitoring, and epidemiological studies. Additionally, it serves as a valuable educational resource for training healthcare professionals in malaria diagnosis. By enhancing the quality and diversity of malaria images, our research contributes to improved diagnosis and treatment strategies. The expanded dataset opens avenues for further research, ultimately leading to enhanced patient care and more effective interventions against malaria.

VII. REFERENCES

Here are the IEEE-style references for the provided links:


[26] Malaria falciparum - Schizont [Photograph]. Center for Young Children’s Health. Retrieved from https://www.cych.org.tw/lab/12-%E6%95%99%E5%AD%B8/autorunDPDx/DPDx/HTML/Frames/M-R/Malaria/falciparum/body_malariadffalcchi.htm


REFERENCES