Malaria Detection using Machine Learning

Harshit Goyal
harshit20203@iiitd.ac.in

Madhava Krishna
madhava20217@iiitd.ac.in

Shreya Bhatia
shreya20542@iiitd.ac.in

Srishti Singh
srishti20409@iiitd.ac.in

Abstract

Malaria is a life-threatening spread by infected Anopheles mosquito bites. Existing means of diagnosis include light microscopy and rapid diagnostic tests, which are used in conjuction to provide accurate results. However, the costs associated with them, in terms of human capital and time required, are immense.

We seek to provide a complementing approach to infection classification using machine learning, which is fast and inexpensive. By training different algorithms like logistic regression, boosted decision trees, support vector machines and convolutional neural networks on images of varying sizes and using image transformations to augment the dataset, we conduct a comprehensive study on model accuracy and inferencing time.

1. Introduction

Malaria is an infectious disease caused by 5 species of the Plasmodium parasite: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale* and *Plasmodium knowlesi*, spread by bites of the Anopheles mosquito. An estimated 241 million infections and 627,000 deaths occurred in 2020-21 [1].

1.1. Testing Methods

The infection can be detected using microscopy tests, Rapid Diagnostic Tests (RDTs) and serological tests.[2]

Microscopy tests involve collecting and dyeing a thin or thick blood specimen with Giemsa or Wright's stain to detect infections visually and ascertain the percentage of infected to uninfected cells.

RDTs indicate whether the patient is infected with one of the species of the malaria-causing *Plasmodium* and provide results in about 15 minutes. However, they fail to indicate a premature infection and negative RDT results need further evaluation. Using microscopy is also advised with positive results, so that the proportion of parasitized to uninfected cells can be determined.

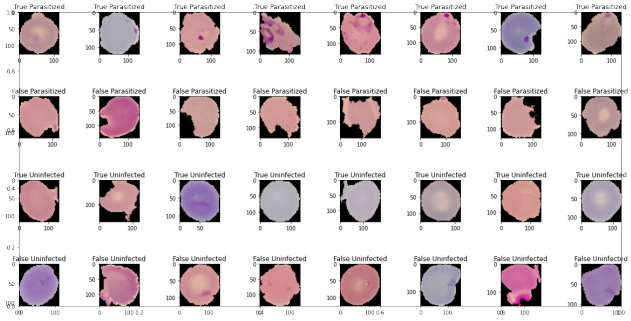

Figure 1. True and false parasitized and uninfected images.

Serological tests examine whether antibodies for the infection are present. They are mostly used for screening blood donors, testing for questionable diagnosis accompanied with treatment.

2. Literature Survey

3. Methodology

4. Dataset

The dataset used was publicly available, courtesy of images were taken at Chittagong Medical College Hospital, Bangladesh.[3]

4.1. Dataset Description

The dataset contains 13,799 parasitized and 13,799 uninfected image samples containing 3 colour dimensions for a total of 27,588 images. The images are of varying sizes. The maximum height and width was 385 and 394 pixels respectively. The minimum height are width was 40 and 46 pixels respectively. The mean height and width was 133 and 132 pixels respectively. The median height and width was 130 pixels. The mean aspect ratio of the images is 1.0138.

Out of the 27,588 images, 647 parasitized and 750 unparasitized images were misclassified [4].

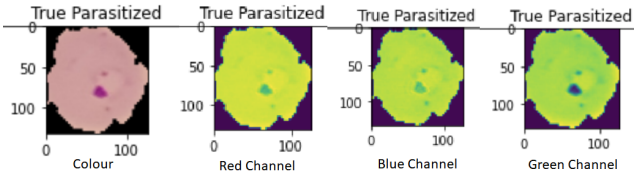

Figure 2. Comparison between various colour channels for a true parasitized cell.

4.2. Exploratory Data Analysis

In order to determine which colour channel was the clearest with respect to the identification of the chromatin dot characteristic to the parasitized cell, we plotted the images in different colour channels and in grayscale.

Out of the plotted images, the green channel showed the maximum isolation of the chromatin dot. We also visualised inverted images and noticed that the green channel had isolated the chromatin dot the most.

4.3. t-SNE visualisation

To visualise the images, we used the t-SNE algorithm to reduce dimensions

4.4. Language

All manuscripts must be in English.

4.5. Dual submission

Please refer to the author guidelines on the CVPR 2020 web page for a discussion of the policy on dual submissions.

4.6. Paper length

Papers, excluding the references section, must be no longer than eight pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of eight pages with two pages of references would have a total length of 10 pages. **There will be no extra page charges for CVPR 2020.**

Overlength papers will simply not be reviewed. This includes papers where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LaTeX guide already sets figure captions and references in a smaller font. The reason such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The reviewing process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in eleven.

4.7. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\cvprfinalcopy` command in the document preamble.) Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g. this line is 095.5), although in most cases one would expect that the approximate location will be adequate.

4.8. Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn't refer to it in the text doesn't mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: `http://www.pamitc.org/documents/mermin.pdf`.

4.9. Blind review

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one's own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

Blind review means that you do not use the words "my" or "our" when citing previous work. That is all. (But see below for techreports.)

Saying "this builds on the work of Lucy Smith [1]" does not say that you are Lucy Smith; it says that you are building on her work. If you are Smith and Jones, do not say "as we show in [7]", say "as Smith and Jones show in [7]" and at the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper just asking to be rejected:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of our previous paper [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.

[1] Removed for blind review

An example of an acceptable paper:

> An analysis of the frobnicatable foo filter.
>
> In this paper we present a performance analysis of the paper of Smith *et al*. [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.
>
> [1] Smith, L and Jones, C. "The frobnicatable foo filter, a fundamental contribution to human knowledge". Nature 381(12), 1-213.

If you are making a submission to another conference at the same time, which covers similar or overlapping material, you may need to refer to that submission in order to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission [2] as additional material and cite it as

> [1] Authors. "The frobnicatable foo filter", F&G 2014 Submission ID 324, Supplied as additional material `fg324.pdf`.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a techreport for further details. Thus, you may say in the body of the paper "further details may be found in [5]". Then submit the techreport as additional material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool which is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties", by Zeus *et al*.

You can handle this paper like any other. Don't write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] didn't handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ

Q: Are acknowledgements OK?
A: No. Leave them for the final copy.

Q: How do I cite my results reported in open challenges?
A: To conform with the double blind review policy, you can report results of other challenge participants together with your results in your paper. For your results, however, you should not identify yourself and should not mention your participation in the challenge. Instead present your results referring to the method proposed in your paper and draw conclusions based on the experimental comparison to other results.

4.10. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TEXbook, p165.

The space after *e.g*., meaning "for example", should not be a sentence-ending space. So *e.g*. is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al*." (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [6], and subsequently developed by Alpher and Fotheringham-Smythe [7], and Alpher *et al*. [8]."

This is incorrect: "... subsequently developed by Alpher *et al*. [7] ..." because reference [7] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*.

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [7, 6, 2] to [6, 7, 2].

5. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

5.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high. Page numbers should be in footer with page numbers, centered and .75 inches from the bottom of the page and make it start at the correct page number rather than the 4321 in the example. To do this fine the line (around line 23)

```
%\ifcvprfinal\pagestyle{empty}\fi
\setcounter{page}{4321}
```

where the number 4321 is your assigned starting page.

Make sure the first page is numbered by commenting out the first page being empty on line 46

```
%\thispagestyle{empty}
```

5.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

Figure and table captions should be 9-point Roman type as in Figures **??** and **??**. Short captions should be centred. Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

5.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

5.4. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [2]. Where appropriate, include the name(s) of editors of referenced books.

5.5. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

5.6. Color

Please refer to the author guidelines on the CVPR 2020 web page for a discussion of the use of color in your document.

6. Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings.

Please direct any questions to the production editor in charge of these proceedings at the IEEE Computer Society Press: `https://www.computer.org/about/contact`.

References

[1] W. H. Organization, *World malaria report 2021*. World Health Organization, 2021. p. 23 Death and Infection Statistics.

[2] C. for Disease Control and Prevention, "Cdc - malaria - diagnostic tools," 2020. Malaria Testing.

[3] N. L. of Medicine, "Thin smear single-cell dataset." Link 1; Link 2; Link 3. Dataset.

[4] F. KMF, T. JF, S. MRA, M. S, M. N, and R. T., "Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application," *Diagnostics (Basel)*, vol. 10, no. 5, p. 329, 2020. Misclassified images.

[5] Authors, "Frobnication tutorial," 2014. Supplied as additional material `tr.pdf`.

[6] F. Alpher, "Frobnication," *Journal of Foo*, vol. 12, no. 1, pp. 234–778, 2002.

[7] F. Alpher and F. Fotheringham-Smythe, "Frobnication revisited," *Journal of Foo*, vol. 13, no. 1, pp. 234–778, 2003.

[8] F. Alpher, F. Fotheringham-Smythe, and F. Gamow, "Can a machine frobnicate?," *Journal of Foo*, vol. 14, no. 1, pp. 234–778, 2004.