Human Malaria Detection and Stage Classification using Random Forest Classifier

Sushil Kumar Mishra

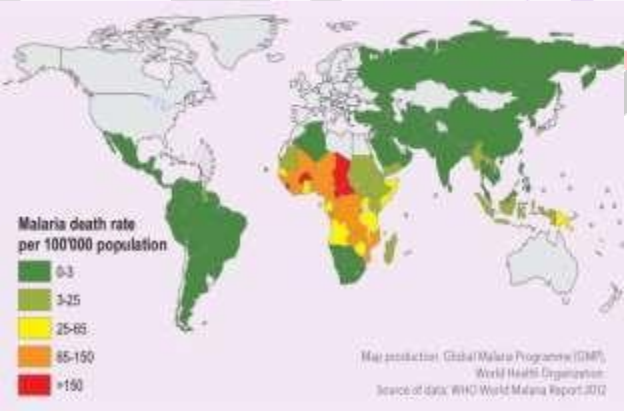
Assistant Professor
Department of Computer Science & Engineering
Chandigarh University

***Abstract*: Malaria is mosquito-borne infection brought about by parasites of the genus Plasmodium. Due to the difficulty of identifying low-abundance parasites from blood serum, early diagnosis of malaria is daunting. Malaria treatment is tested by microscope patient stressed blood stain. The blood sample to be examined is put in a transparent glass slide under a microscope to count the number of RBC contaminated concentrations. To view the slide with extreme visual focus, an intended microscopist is required. This entire process is time- consuming, exhausted and less accurate. Under the roof of this paper, we build a fully automatic detection system to count the plasmodium parasites in blood Smear. This system is based on an algorithm for machine learning against traditional thin-blood smear analysis, which has more sensitivity and specificity. This automated system is simple, making it ideal even with low levels of parasites for ultra-fast and accurate detection. The suggested technique utilizes collected photographs of patients to assess the malaria disease without staining the blood or specialist required.**

***Keywords*: RBC, plasmodium, blood smear, microscopy**

I. INTRODUCTION

Malaria is also referred to as jungle fever. Malaria is a life- threatening disease caused by parasites spreads to humans through the bites of infected female Anopheles mosquitoes, the worm infects the red blood cell. It can be avoided or treated. Malaria is caused by larvae called Plasmodium. Bites of pregnant female Anopheles mosquitoes, or malaria vectors, transmit the disease to humans. Using a microscope to diagnose malaria disease is a time- consuming and difficult process. The traditional method involves the microscopist or laboratory technician's tremendous skills. According to the WHO study 2012, this world map shows the countries with significant death rates from malaria.


Malaria death rate per 100'000 population
0-3
3-25
25-65
65-150
>150
Map production: Global Malaria Programme (GMP), World Health Organization
Source of data: WHO World Malaria Report 2012

Globally, in 92 nations, an estimated 3.4 billion people are at risk of malaria infection or disease and 1.1 billion people are at high risk (> 1 in 1000, malaria chances per year). According to the 2018 World Malaria Survey, there were 219 million cases of malaria worldwide in 2017 (uncertainty estimate 203–262 million) and 435 000 deaths from malaria, and 214 million new cases of malaria were recorded in 2015, resulting in 438,000 deaths, reflecting a drop in malaria cases and death rates of 18% and 28% since 2010, respectively. The burden was more substantial inthe African region of the WHO, where an estimated 93 percent of all malaria deaths occurred, and 61 percent of all deaths in children under the age of 5. Five countries accounted for almost half of all outbreaks of malaria worldwide in 2017: Nigeria (25%), the Democratic Republic of the Congo (11%), Mozambique (5%), India (4%) and Uganda (4%).

Diagnosis of infection is a significant problem in developing countries such as Uganda [6] and many African countries, where only half of the rural health centres have microscopes and almost one-fourth have qualified malaria laboratory technicians. It is also necessary to identify the diseases with greater accuracy at the earliest stage, as it can serve to provide the confirmed person with the treatment at an early stage. In contrast, false negatives could affect the fatality, or false positives may lead to an increase in needless economic burden or drug resistance. Thus, a different method of diagnosis needs to be developed.

Image analysis and automatic testing program can be introduced. We are proposing a new automated system focused on the methodology of the microscopy photos to classify the malarial parasite. For the segment identification of malaria parasite tissue, this project uses a random forest algorithm. Machine learning algorithms have been provided with sufficient training data. The blood smear parasites are identified using standard microscope photographed images. Few other studies further investigated the clustering of the various species and the various stages of the life cycle of the parasites. Image processing methods are still being practised because we don't want to completely wipe out the diagnostic process of human experts, but to some extent for final blood smear judgment. This system would improve the efficiency of laboratory technicians by helping to check their focus and also by introducing malaria treatment through a remote network connection.

This paper focuses on automatic malaria identification by identifying and classifying stable erythrocytes from contaminated erythrocytes in photographs with low quality blood smear. We use highly efficient computer algorithms as these low-quality images are not handled by traditional algorithms. Therefore, without any human intervention, our device can diagnose malaria or at least the system can act as a valuable tool for technicians to minimize their job and also improve diagnostic accuracy.

II. RELATED WORK

Specific thresholding was suggested in [7] to recognize the presence of Plasmodium in the thin smears of the blood. Many malaria detection methodologies are focused on two criteria: (i) photographs obtained under well-controlled conditions; (ii) the need for sufficient microscope equipment. Both criteria are different in the endemic area of malaria, but we cannot entirely rely on these methods in many parts of the world because of less availability of doctors or technicians. We, therefore recommend an automated system using a microscopic image of the infected person's thin blood smear. This uses microscopic photo to examine the amount of parasite concentrations found in the specific RBC on RBCs. The area of RBC parasite disease determine such levels of parasites. The device thus classifies the blood smear obtained as parasitized and contaminated. Parasite-containing patches are marked as positive and the remaining patches are marked as negative. The foundation of this framework is conventional interface design and a decision-tree classifier ensemble. The findings recorded using traditional engineering features (morphology, form, colour and texture).

Our system is designed to be used in the sector. Thus, the device has the following requirements and characteristics:
(1) embraces regular field-prepared Giemsa slides;
(2) is resilient to significant slide value variability;
(3) scans a reasonable blood volume, approximately 0.1 μL, approximately 300 FoVs;
(4) scans on several focal planes;
(5) has a strong patient-level sensitivity and specificity at low parasitemia — approximately 100 p/μL;
(6) has an acceptable parasitemic quantity of 200-200,000 p/μL; and
(7) has a strong sensitivity and specific ity to the sample.

A lot of new methodologies for malaria diagnosis have been established in recent years, including rapid antigen detection, fluorescent microscopy detection method, and PCR (Polymerase Chain Reaction) process for detecting different nucleic acid sequences. Given that, the process for diagnosing light microscopy is the most commonly employed procedure. Microscopy may be used to display the RBC's frequency. The program is taught with a wide and varied set of images where the sample collection is isolated from the patient-level training. We broke the whole dataset into the 99:1 combination for training and testing sets respectively. The system can make choices with an accuracy rate of 91 percent. Our program reaches tier 1 of WHO competence for P. falciparum diagnosis and adequate reliability of P. falciparum quantitation to be used for studies in drug resistance.

III. METHODOLOGY

Our data is diverse, with a lot of image variation. Developing an algorithm which works for a particular image may be straightforward, but creating an algorithm which works across the board of heterogeneous datasets is a very different challenge. There are several solutions that can be used, a versatile complex method that has good diagnosis performance and accuracy.

A. Data Collection

For this model, photographs are the main sources of information. The scientists at the Lister Hill National Center for Biomedical Communications (LHNCBC), part of the National Medicine Library (NLM), took these pictures. At the Chittagong Medical College Hospital in Bangladesh, Giemsa-stained thin smear slides from 150 P. falciparum-infected and 50 stable patients are gathered and documented. A professional slide reader at the Mahidol- Oxford Tropical Medicine Research Unit in Bangkok, Thailand, manually annotated the images. The database contains a total of 27,558 cell objects of equivalent parasitized and uninfected cell instances. CSV file is maintained for the parasitized and uninfected groups comprising the cell mappings of each RBC’s image.

B. Proposed Architecture

The machine learning algorithm takes information from task-related inputs and outputs and constructs a software that can distinguish them automatically. We're not going to discuss machine learning in this study but we're going to show the framework we've used for this work.

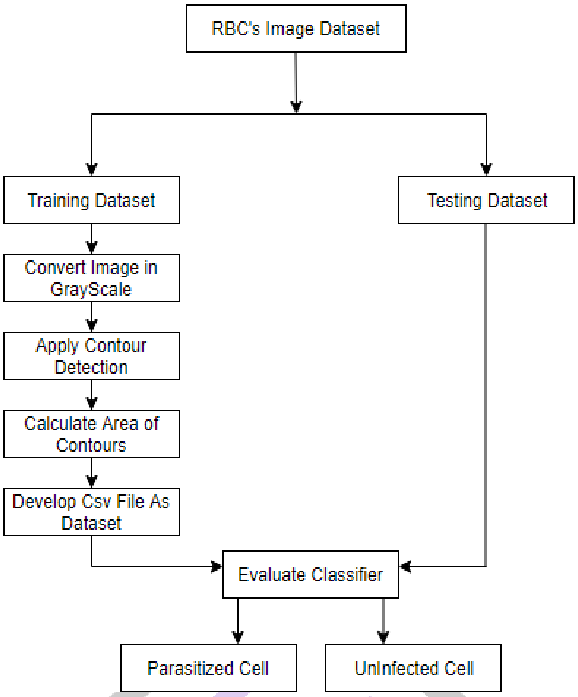


Figure 1: Machine learning algorithm to learn, detect the parasitized cells.

The identification module creates suggestions for objects— potential parasites that are ultimately graded by a classifier as parasites or distractors. Because of the Poisson statistics for uncommon object distributions, some~300 FoVs need to be processed by the algorithm to reach the goal detection maximum. Many standard methods of object detection, such as R-CNN[9], YoLo[10], deformable pieces model[11], and selective scan are either too complicated, too naive, or too sluggish to detect malaria on several focal planes in large numbers of FoVs.

While this basic detector has high sensitivity, its reliability is poor: several dark distractors are also identified, which due to excessive false positive detections degrades low parasitic efficiency. We are adding two developments to improve the detector's target level specificity: adaptive grayscale frequency and flexible spatial thresholding. The standard grayscale brightness is a linear mixture of red, green, and blue pixel values that approximates the luminance observed by humans, but does not generally provide the optimal parasite-background isolation. A more efficient prediction matrix can be determined using machine learning techniques.

Every 1024 x 768 image labelled as either uninfected or parasitized was divided into patches overlapping. Every width of the picture patch is 50-50 pixels. We address malaria detection function as a binary classification method with the marked object patches dataset. In the identification function, the raw encoded pixel information for object patches will not be directly useful. Instead, they use a model that will not be influenced by the speed of encoding, rotation, or continuous offsets. In the Plasmodium detection issue, the form of artefacts in the output patches is the main concern.

An essential step in the development of the automated malaria diagnostic system is feature engineering First we seek a representation of the data resulting in good performance on plasmodium detection and then have a general representation of the shapes found in the images comprising blood smear including artefacts such as leukocytes or the numerous hemiparasites, so that the same method can be used in the future to classify the other related problems. Usually, Colour data can also be very helpful, although it is not insightful when using blood films that are contaminated with the colour of the ground. For this task statistical representations of the shapes are used. Normally, for extraction of the functionality, we need to convert the colour patches to grayscale patches.

Because of these labelled image patches, the detection of malaria may be posed as a classification problem, i.e. classification of either 0 (uninfected) or 1 (parasitized). We use Random Forest Machine Learning algorithms because even with 0.907 precision, it is very good to diagnose malaria. The quality was assessed in the current work focused on the existence of parasites at the patch level and not at that patient's whole picture stage. The patient is considered to be infected if the image specimen includes at least a positive patch. Since the photographs we have for our tests are from malaria-infected individuals, per-patient tolerance and specificity results cannot be provided.

This program can be used as a support system, making it easy for technicians to make the decision. This helps in analyzing the images taken from the microscope to concentrate the expert's attention on the artefacts within those pictures that are more likely to contain Plasmodium. A specific threshold of greater sensitivity is chosen for this reason. We utilize different classification criteria to achieve specific false positives and negatives.

This system uses machine learning package OpenCV for contour detection to detect parasitized patches of images.

How to draw the contours?

Cv.drawContours function is used to draw the contours. It can also be used to depict any shape as long as you have the boundaries. The first argument is the source image; the second argument is the contours that should be passed as a Python list, the third argument is the contours index (useful when drawing individual contours, pass-1) and the remaining arguments are colour, thickness, etc.

To display all the contours in an image:

```
cnt = contours[4]
cv.drawContours(img, [cnt], 0, (0,255,0), 3)
```

This system Uses Python3.7 with Sci-Kit Learn [10] and OpenCV2 [11] to implement. This experiment was done on a CPU system with the installation of the 8 GB RAM and i5 processor.

IV. RESULTS AND DISCUSSION

This configured and pre-trained model's efficiency is examined to the task of classifying parasitized and uninfected cells. The model is trained and optimized to minimize cross-entropic loss and categorize the images of cells to their classes. Whenever the accuracy of the validation has ceased to improve, the learning rate is reduced. The failures in training and testing reduced with epochs suggesting the learning process. An ensemble system produces positive results only when the individual base-learners have sufficient heterogeneity.

	Precision	Recall	F1-score	Support
Parasitized	0.90	0.92	0.91	676
Uninfected	0.92	0.90	0.91	701
Accuracy			0.91	1377
Macro avg	0.91	0.91	0.91	1377
Weighted avg	0.91	0.91	0.91	1377

Table 1: Performance metrics of the proposed model using Random forest, machine learning algorithm.

Random Forest outperformed every other algorithm out of any Machine learning algorithms, so we used random forest classification. The precision, accuracy, recall, and f- score values are tabulated in Table 1. Some terminology used in the table here is like fellows.

True Positives (TP) — These are the accurately estimated positive values, meaning the actual class value is yes and the projected class value is yes as well.

True Negatives (TN) — These are the accurately estimated negative values indicating that the actual major value is no and the expected class value is no as well.

False Positives (FP) — When the actual class is no and the predicted classification is also no.

False Negatives (FN) –If the real category is true, but no class is expected.

Accuracy = TP+TN/TP+FP+FN+TN Accuracy= 0.91

Precision = TP/TP+FP Precision= 0.90 ± 0.02

Recall = TP/TP+FN Recall = 0.90 ± 0.02

F1 Score = 2*(Recall * Precision) / (Recall + Precision) F1 Score = 0.91

We ensured that the Random Forest offers an optimal solution by (a) optimization of design and hyper parameter,

(b) implied regularization enforced by batch standardization, and (c) enhanced generalization by violent dropouts in convolutional and dense layers. At the patient stage, we performed cross-validation to provide a real quality assessment for predictive models so that the test data reflects truly unobserved objects without leakage of knowledge from the training data related to the staining differences or other items. This model, however is crucial in selecting diversified base-learners in the function space that is appropriate in various regions. For this purpose, for the current mission, we decided the best model combination to develop the ensemble. Experimental results are statistically significant if they are not due to chance and a relationship actually occurs for a given level of statistical significance. We carried out quantitative analyses to assess the presence of a statistically significant variation in the quality of the models being examined.

CONCLUSION

It is seen that model ensemble utilizing multiple models of Machine Learning produced promising predictive output that none of the individual constituent models can achieve. Ensemble learning decreases model uncertainty by optimally integrating multi-model forecasts and increasing exposure to training data and algorithms details. Through integrating the ensemble paradigm into a system, they have built a web application to mitigate privacy issues, low latency, and provide cross-platform benefits. The model ensemble's output simulates real-world conditions with minimal uncertainty, overfitting, and leads to better generalization. We assume the proposed findings will be helpful in designing medically approaches for the identification and separation of parasitized and uninfected cells in images of thin-blood smear.

V. ACKNOWLEDGEMENTS

We want to thank the researchers at the Lister Hill National Center for Biomedical Communications (LHNCBC) and National Library of Medicine (NLM) to grant permission to use data and material required for our research. We are indebted to our University and our professor Sushil Kumar Mishra for her expert guidance to enhance our research.

REFERENCES

[1] https://www.researchgate.net/publication/322819026_Malaria_Detection_Using_Image_Processing_and_Machine_Learning
[2] https://pubs.rsc.org/en/content/articlelanding/2014/l c/c4lc01058b/unauth#!divAbstract
[3] https://www.isglobal.org/en/-/malaria-eradication- research-agenda-malera-initiative
[4] World Health Organization (2010) World Malaria Report. Geneva: World Health Organization.
[5] World Health Organization (2012) Disease surveillance for malaria control. Geneva: World Health Organization. (In press).
[6] Quinn JA, Andama A, Munabi I, Kiwanuka FN, Automated blood smear analysis for mobile malaria diagnosis, Mobile Pointof-Care Monitors and Diagnostic Device Design, (2014), 31-115
Pointof-Care Monitors and Diagnostic Device Design, (2014), 31-115.
[7] Tumwebaze M, Evaluation Of The Capacity To Appropriately Diagnose And Treat Malaria At Rural Health Centers In Kabarole District, Western Uganda, health policy and development, 9(2011),46- 51 detection. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779- 788, 2016.
[8] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Raman. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1627-1645, 2010.
[9] https://blog.exsilio.com/all/accuracy-precision- recall-f1-score-interpretation-of-performance- measures/
[10] https://towardsdatascience.com/detecting-malaria- with-deep-learning-9e45c1e34b60
[11] https://www.sciencedirect.com/science/article/pii/S 193152441730333X
[12] https://www.sciencedirect.com/science/article/abs/p ii/S0001706X03003164
[13] https://www.ajtmh.org/content/journals/10.4269/ajt mh.1999.60.687
[14] https://onlinelibrary.wiley.com/doi/full/10.1046/j.13 65-2141.1999.01199.x
[15] https://pubs.rsc.org/en/content/articlelanding/2014/l c/c4lc01058b/unauth#!divAbstract
[16] https://journals.plos.org/plosmedicine/article?id=10. 1371/journal.pmed.1001142#s1
[17] https://peerj.com/articles/6977/#materials|methods
[18] Anggraini D, Nugroho AS, Pratama C, Rozi IE, Pragesjvara V, Gunawan M, Automated status identification of microscopic images obtained from malaria thin blood smears using Bayes decision: a study case in Plasmodium falciparum, International Conference on Advanced Computer Science and Information System (ICACSIS), (2011), 347-352.
[19] https://lhncbc.nlm.nih.gov/publication/pub9932
[20] R.B. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580- 587, 2014.
[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object.
[22] Basic Malaria Microscopy: Tutor's guide, WHO(World Health Organization, (2010).
[23] Di Ruberto C, Dempster A, Khan S, Jarra B, Automatic thresholding of infected blood images using granulometry and regional extrema, Pattern Recognition,3(2000), 441-444.
[24] Di Ruberto C, Dempster A, Khan S, Jarra B, Analysis of infected blood cell images using morphological operators, Image and vision computing, 20(2002),133-146.
[25] Samba EM, The burden of malaria in Africa, Africa health, 19(1997), 17.
[26] CDC, "Frequently asked questions (FAQs)," CDC, 2016. [Online].
Available: https://www.cdc.gov/malaria/about/faqs.html. Accessed: Jan. 25, 2017.
[27] D Ghate, C. Jadhav, and N. U. Rani, "AUTOMATIC DETECTION OF MALARIA PARASITE FROM BLOOD IMAGES,". [Online]. Available: http://ijact.org/volume4issue1/IJ0410050.pdf. Accessed: Jan. 25, 2017.
[28] Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW (2004) The global distribution and population at risk of malaria: past, present and future. Lancet Infect Dis 4: 327–336.
[29] D.M. Memeu, K.A. Kaduki, A. Mjomba, N.S. Muri uki, L. Gitonga Detection of plasmodium parasites from images of thin blood smears
Open J Clin Diagnostics, 3 (2013), p. 183