PERFORMANCE COMPARISON OF MACHINE LEARNING ALGORITHMS FOR MALARIA DETECTION USING MICROSCOPIC IMAGES

Saiprasath G, Naren Babu R, ArunPriyan J, Vinayakumar R, Sowmya V,
Soman K P
Centre for Computational Engineering and Networking (CEN),
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidhyapeetham, India

***Abstract*:** Malaria is a blood-borne disease by mosquito caused by Plasmodium parasites. The standard method for malaria detection involves preparing a blood smear and examining the stained blood smear using a microscope to detect the parasite genus Plasmodium, which heavily relies on the expertise of trained experts. Under the roof of this paper, with the intention of singling out the parasite blood smears for malaria detection, shallow machine learning algorithms are used against the traditional method, which has some snags related to sensitivity and specificity. The proposed methodology determines the malarial infection with the help of captured images of patients without staining the blood or need of experts.

Key Words: **Microscopy, Malaria, Plasmodium, blood smears, Machine Learning, Decision Trees, Naive Bayes, Random Forest Tree, Adaboost, Logistic Regression**

1. INTRODUCTION

A survey by WHO (World Health Organization) predicted that malaria could occur nearly in 33 hundred million cases [1]. Blood-borne disease, Malaria is caused by Plasmodium, red blood cells infected by the parasite and it is spread by the specific kind of mosquito called Anopheles. A person affected by malaria will show many clinical manifestations from very mild to severe cases, which might even lead to the death of the person. The detection of malaria disease using a microscope is a time-consuming and difficult process. This traditional method needs the considerable expertise of microscopist or laboratory technician. Experienced malaria microscopist indeed plays an important role in parasite identification [2], [3]. According to the research conducted in [5], [8] and [13], it is reported that 1-3 million are nearly fatal out of 300-500 million of acute malarial diseases. In regions major affected by malaria, diagnosis is very difficult and treatments are given based on symptoms alone.

Diagnosing disease is a major problem in developing countries like Uganda [14], where only half of the rural health centres have microscopes and nearly only one-fourth of them have trained laboratory technicians for malaria diagnosis. Also, detecting the diseases at the earliest with better accuracy is important, as it may help in providing the medication to the diagnosed patient at an early stage. Moreover, the fatality could be caused by false negatives, and false positives could cause increase in unnecessary economic burden and drug resistance [21], [16]**.** Therefore, there is a need to develop a different method for diagnosis.

Image processing and the computer vision methods can be implemented for diagnosis. Recently, Khan and his dedicated team proposed a new computer vision method based on the approach to identify the MP (Malarial Parasite) from the light microscopy images. This is a pixel-based approach, which uses K-means clustering algorithm for the segment identification of malaria parasite tissue [22]. In [4], enough training data were provided to the machine learning algorithms. The parasites present in blood smear are identified using images photographed via standard microscope. Few other studies looked even further at clustering the different species and the different stages of the parasites life cycle [23], [25]. Image processing methods are still being practiced because, we don't want to wipe out human experts diagnostic process completely, but in a certain degree for final judgment based on the blood smear. This process will improve the efficacy of lab technicians by helping to triage their concentration and also implement the malaria diagnosis over a remote network connection.

This paper focuses on automated malaria detection by localizing and classifying the infected erythrocytes from healthy ones in low quality blood smear images. We use classical machine algorithms because, conventional algorithms fail to process these low-quality images. Thus, our system can detect malaria without any human intervention or at least the system can serve as a helpful medium for technicians to reduce their work and also possibly increment the diagnostic accuracy.

The rest of the sections are organized accordingly: section 2 briefly discusses and analyses the existing works in Malaria diagnosis and the standard practices followed. Employing the ML (Machine Learning) classifiers to carry out the diagnosis of Malaria as an object detection model and also the methods used for extracting the statistical image features are discussed in section 3. Experimental analysis and results of our system are presented in section 4 followed by conclusion of our paper in section 5 and the possible future directions are explained in section 6.

2. RELATED WORKS

Malaria is bred by the parasite Plasmodium, which attacks red blood cells (RBC) and is transfused by mosquitoes. Malaria's severity ranges mild to highly serious, which eventually leads to the death of humans. Neural networks have been used in analyzing the possibility of RBC's and parasite in the blood smear [24]. In [19], the weighted KNN (K-Nearest Neighbors) algorithm is trained with the learned features using the Bayesian pixel classifier, whose purpose is staining pixels. For identifying multi-class parasites in terms of lifecycle stage and its types are attempted in [15]. Basic thresholding is done using a histogram based method to identify the existence of Plasmodium in the blood smears is proposed in [9]. Smear preparation is important, as differences in these might cause variations, as predominant as imaging conditions [7].

The overlapping RBC's were separated using morphological operators [10], [11]. The abnormalities are reported by analyzing of blood cell images, where the true image is binarized applying a fuzzy measure technique and then cells present in the image are labelled [6]. Further, these labelled cells are classified into platelets, leukocyte, and erythrocyte using a architecture called hierarchical Neural Network using some attributes like color, size, and features. In [20], the algorithm consists of four stages i.e detecting the edges, linking the edges, clump splitting the clumps and then detecting the Parasite. Pre-processing used in this algorithm is adaptive histogram equalization. A color segmentation technique is used in [20] for separating the pixels into erythrocyte, parasites, and the background, which is based on classical supervised classification models. Supervised classification algorithms like Support Vector Machines, K-Nearest Neighbour, and Naive Bayes were evaluated using different color models namely RGB model (Red Green Blue), normalized RGB model, HSV model (Hue Saturation Value), and YCbCr model respectively.

There has been a great deal of developing new methodologies in last few years for malaria diagnosis, which includes rapid antigen, fluorescent microscopy detection method, and PCR(Polymerase Chain Reaction) method that detect the specific sequences of nucleic acid [17]. In spite of this, light microscopy diagnosis method is the most widely and commonly used technique [18]. In [29], edges are detected using the sparse banded filter matrices and in [28], the tumor is classified using X-ray image classification. Microscopy can differentiate between the types of species, quantify parasitemia and examine the different asexual stages of the parasite [18] and [20]. But, this technique needs trained technician and it's a time- consuming process and the final preciseness of the diagnosis is dependent on experience and skills of the microscopist and also the amount of time invested in learning each slide [20].

3. PROPOSED METHODOLOGY

The image dataset used in this paper is captured using oil immersion objective lens from 133 individuals using with 1000x magnification [4] (http://air.ug/downloads/plasmodium-images.zip). The images which were out-of-focus, poor quality images and images which were difficult to label the parasite were eliminated. Finally the dataset contains 2703 blood smear images with bounding boxes of 50,255 of malaria parasites. Then, each image is split into overlapping patches and the patches are labelled as 0 or 1 using the bounding box. Training dataset uses 75% of the labelled data i.e. 2027 blood smear images, which contains 37550 patches where it is annotated as Parasites, and the remaining 25% of the data is used for testing purpose i.e. 676 blood smear images, which contains 16312 patches annotated as Parasites.

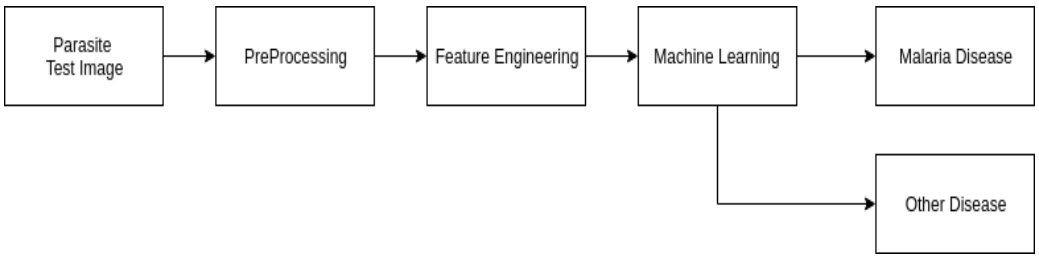

Figure 1: Proposed Framework for applying machine learning algorithms for malaria detection using microscopic images

Figure 1 shows that the test image of the patient is pre-processed and feature engineered before feeding to the Machine Learning algorithm. Then, the binary classification algorithm classifies the image patch as presence or absence of Malaria.

Each 1024×768 image which was labelled as either 0 or 1 was split into nearly 475 overlapping patches. Each image patch size is of 50×50 pixels. Having this labelled image patches dataset, we pose malaria identification task as a binary classification model. The raw formatted pixel data in image patches won't be directly useful in the classification task. Instead, we use a representation, which won't be affected by translation, rotation, and constant offsets in the intensity. The shape of objects in the input patches is the main concern in the Plasmodium detection problem. We need to scale the images if they are collected with different sizes and require a representation invariant to translation, intensity, and rotation.

Feature engineering is an essential step in developing the automatic malaria diagnosis system. First, we need to find a representation of the data which results in good performance on detecting the plasmodium, and then to have a general representation of the shapes present in the images containing blood smear excluding objects such as leukocyte or the different hemoparasites, so that in future the same platform can be used to identify the other related problems. Generally, color information can also be very useful, though it isn't informative when using blood films, which are treated with the field's stain. Therefore, statistical representations of the shapes are used for this task. Generally, we need to convert the color patches to grayscale patches for feature extraction. Here two types of features are being used: one is derived from connected components, and other is derived by calculating the moment's patches, thresholded at multiple levels.

Given this labeled image patches, the Malaria detection can be posed as a classification problem i.e. classifying either 0 (another disease) or 1(Malaria disease). We use several Machine Learning algorithms such as AdaBoost, Random Forest, Decision Tree, KNN to detect the malaria. Random Forest was very good in good in detecting malaria with accuracy of 0.965. In the current work, the performance is evaluated based on the presence of parasite at the patch level and not at the entire image level of each patient. The person is declared infected; if there exist at least on a positive patch in the images sample. Since the images we have for our experiments were from the individuals infected with malaria, it's not possible to give per-patient sensitivity and specificity results.

This system can be used as a helping system, so technicians can make the decision easily. This results in processing the images taken from the microscope for making the expert's attention to focus only on the objects within those images that are more likely of the containing Plasmodium. For this purpose, a different threshold is selected with greater sensitivity. For getting different false positives and negatives we use different classification thresholds.

The implementation of this system is done using Python2 with Sci-Kit Learn [26] and OpenCV2 [27]. This experiment was performed on a CPU system with the 32 GB RAM and i7 processor configuration.

4. EXPERIMENTAL ANALYSIS AND RESULTS

The image dataset used in this paper is obtained from [4]. Training dataset uses 75% of the labeled data i.e 2027 blood smear images, which contains 37550 patches, where it is annotated as Parasites, and the remaining 25% of the data is used for testing purpose i.e 676 blood smear images, which contains 16312 patches annotated as Parasites.

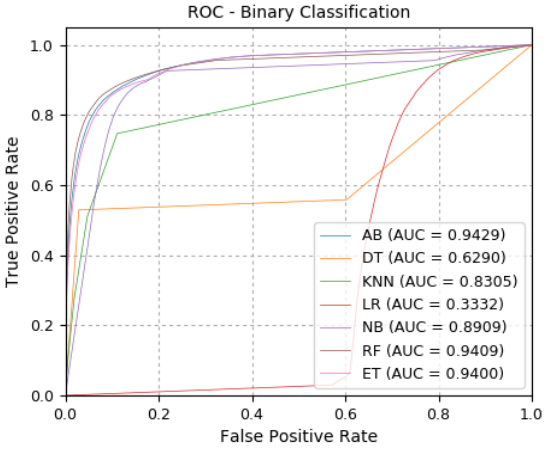

Fig 2. ROC Curve

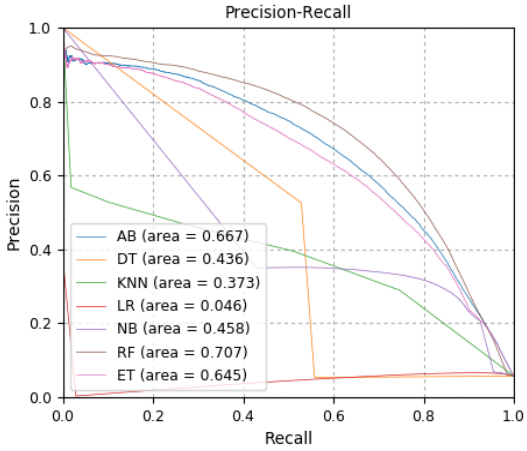

Fig 3. Precision-Recall Curve

Fig. 2 represents ROC (Receiver Operating Characteristics) and it can be seen that the AUC (Area Under the Curve) to be 0.943 for AdaBoost, which shows its capacity in distinguishing the positive and negative patches.

Precision – Recall (PR) Curve, which is graphed in Fig. 3, shows the possible trade-offs between decreasing the false alarm rate and increasing sensitivity. If the 90% precision is chosen as detection threshold, the corresponding 20% recall would be higher than any method used to examine thin blood smears using the same no. of fields of view. Fig. 3 concludes that, if we are using this detection system for completely automated malaria diagnosis, so that we could accomplish the false alarm rate below 1/10, and the recall value would be around 1/5 of what a trained technician would be able to achieve i.e. nearly 5x higher for the automated diagnosis system compared with the human technician.

Table 1 Performance Measure of the proposed work for malaria detection using classical machine learning algorithms

Algorithm	Accuracy	Precision	Recall	F-Score
Ada Boost	0.962	0.729	0.526	0.611
Decision Tree	0.946	0.526	0.529	0.527
KNN	0.940	0.465	0.278	0.348
Linear Regression	0.943	0.375	0.000	0.000

Naive Bayes	0.858	0.271	0.873	0.414
Random Forest	0.965	0.775	0.553	0.645
Extra Tress	0.956	0.837	0.298	0.440

The performance of the different machine learning algorithms is tabulated in the above table 1. The values of Accuracy, Precision, Recall, F-Score are tabulated. Out of 7 classical Machine learning algorithms we used, Random Forest outperformed every other algorithm closely followed by Ada Boost.

5.CONCLUSION

We have proposed a Malaria parasite detection method using a shallow machine learning algorithms. This method of detecting the malaria parasite can be very useful to health workers in countries, where there is less number of trained laboratory experts and lack of resources. In the present work, we divided the image into patches and analysed based on the presence or absence of malarial parasite. To accomplish this, we have used various classical machine learning algorithms such as AdaBoost, Decision Tree, KNN, Random Forest, etc.

The accuracy of our model assists the laboratory technicians in decision making, by focussing on the parts of images prone to Plasmodium. Automated Malaria diagnosis also aids for data collection. Furthermore, our extraction of features and classification framework may be sufficiently general for other diagnostic tests like hemoparasites, worm infestations, or tuberculosis.

6. FUTURE WORK

Our platform for the automatic diagnosis of Malaria provides many useful and interesting directions in this area. Also, Deep Learning methods can be applied, where accuracy can be pretty much higher than our shallow machine learning methods. Also, adding more layers to the Neural Network can boost up the accuracy.

7.REFERENCES

[1] W. H. O. (World H. Organization), *Global Health Observatory - Malaria*, (2011), Available: http://www.who.int/gho/malaria.

[2] Mehrjou A, Abbasian T, Izadi M, Automatic malaria diagnosis system, *International Conference on Robotics and Mechatronics(ICRoM)*, (2013), 205-211.

[3] McKenzie FE, Sirichaisinthop J, Miller RS, Gasser Jr RA, Wongsrichanalai C, Dependence of malaria detection and species diagnosis by microscopy on parasite density, *The American journal of tropical medicine and hygiene*, **69**(2003), 372-376.

[4] Quinn JA, Andama A, Munabi I, Kiwanuka FN, Automated blood smear analysis for mobile malaria diagnosis, *Mobile Point-of-Care Monitors and Diagnostic Device Design*, (2014), 31-115.

[5] Rafael ME, Taylor T, Magill A, Lim YW, Girosi F, Allan R, Reducing the burden of childhood malaria in Africa: the role of improved, *Nature*, (2006), 39-48.

[6] Kim KS, Kim PK, Song JJ, Park YC, Analyzing blood cell image to distinguish its abnormalities (poster session), *ACM international conference on Multimedia*, (2000), 395-397.

[7] Basic Malaria Microscopy: Tutor's guide, *WHO(World Health Organization*, (2010).

[8] Roca-Feltrer A, Carneiro I, Armstrong Schellenberg JR, Estimates of the burden of malaria morbidity in Africa in children under the age of 5 years, *Tropical medicine & international health*, **13**(2008), 771-783.

[9] 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.

[10] 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.

[11] 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.

[12] Díaz G, Gonzalez F, Romero E, Automatic clump splitting for cell quantification in microscopical images, *Iberoamerican Congress on Pattern Recognition*, (2007), 763-772.

[13] Samba EM, The burden of malaria in Africa, *Africa health*, **19**(1997), 17.

[14] 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.

[15] Tek FB, Dempster AG, Kale İ, Parasite detection and identification for automated thin blood film malaria diagnosis, *Computer vision and image understanding*,**11**(2010), 21-32.

[16]Lee JH, Jang JW, Cho CH, Kim JY, Han ET, Yun SG, Lim CS, False-positive results for rapid diagnostic tests for malaria in patients with rheumatoid factor, *Journal of clinical microbiology*, **52**(2014), 3784-3787.

[17] Haditsch M, Quality and reliability of current malaria diagnostic methods, *Travel medicine and infectious disease*,**2**(2004), 149-160.

[18] Thung F, Suwardi IS, Blood parasite identification using feature based recognition, *International Conference on Electrical Engineering and Informatics*, (2011), 1-4.

[19] Tek FB, Dempster AG, Kale I, Malaria Parasite Detection in Peripheral Blood Images, *British Machine Vision Conference*, (2006), 347-356.

[20] Pammenter MD, Techniques for the diagnosis of malaria, *South African medical journal= Suid-Afrikaanse tydskrif vir geneeskunde*,**74**(1988), 55-57.

[21] Agnihotri N, Agnihotri A, Wrong Sample Dispensing May Cause False Positive Malaria Test, *Journal of Clinical and Diagnostic Research*, 9(2015), EG01-EG02.

[22] Khan NA, Pervaz H, Latif AK, Musharraf A, Unsupervised identification of malaria parasites using computer vision, *International Joint Conference onComputer Science and Software Engineering*, (2014), 263-267.

[23] Lom J, Dyková I, Myxozoan genera: definition and notes on taxonomy, life-cycle terminology and pathogenic species, *Folia parasitologica*,**53**(2006), 1-36.

[24] Ross NE, Pritchard CJ, Rubin DM, Duse AG, Automated image processing method for the diagnosis and classification of malaria on thin blood smears, *Medical and Biological Engineering and Computing*,**44**(2006),427-436.

[25] Dobson A, Lafferty KD, Kuris AM, Hechinger RF, Jetz W, Homage to Linnaeus: how many parasites? How many hosts?, *National Academy of Sciences*,**105**(2008), 11482-11489.

[26] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Scikit-learn: Machine learning in Python, *Journal of machine learning research*,**12**(2011), 2825-2830.

[27] Bradski G, Kaehler A, Learning OpenCV: Computer vision with the OpenCV library, *O'Reilly Media, Inc.*, (2008).

[28] Pooja A, Mamtha R, Sowmya V, Soman K. P, X-ray image classification based on tumor using GURLS and LIBSVM, *International Conference on Communication and Signal Processing*, (2016), 521-524.

[29] Sowmya V, Mohan N, & Soman K. P, Edge detection using sparse banded filter matrices, *Procedia Computer Science*, ***58***(2015), 10-17.