Malaria Detection and Classification Using Machine Learning Algorithms

Yaecob Girmay Gezahegn[1(✉)], Yirga Hagos G. Medhin[2], Eneyew Adugna Etsub[1], and Gereziher Niguse G. Tekele[2]

[1] Addis Ababa University, Addis Ababa, Ethiopia
yaecob.girmay@gmail.com, eneyew_a@yahoo.com
[2] Mekelle University, Mekelle, Ethiopia
yirgaatmail@gmail.com, gezubashay@gmail.com

Abstract. Malaria is one of the most infectious diseases, specifically in tropical areas where it affects millions of lives each year. Manual laboratory diagnosis of Malaria needs careful examination to distinguish infected and healthy Red Blood Cells (RBCs). However, it is time consuming, needs experience, and may face inaccurate lab results due to human errors. As a result, doctors and specialists are likely to provide improper prescriptions. With the current technological advancement, the whole diagnosis process can be automated. Hence, automating the process needs analysis of the infected blood smear images so as to provide reliable, objective result, rapid, accurate, low cost and easily interpretable outcome. In this paper comparison of conventional image segmentation techniques for extracting Malaria infected RBC are presented. In addition, Scale Invariant Feature Transform (SIFT) for extraction of features and Support Vector Machine (SVM) for classification are also discussed. SVM is used to classify the features which are extracted using SIFT. The overall performance measures of the experimentation are, accuracy (78.89%), sensitivity (80%) and specificity (76.67%). As the dataset used for training and testing is increased, the performance measures can also be increased. This technique facilitates and translates microscopy diagnosis of Malaria to a computer platform so that reliability of the treatment and lack of medical expertise can be solved wherever the technique is employed.

Keywords: Machine learning · Image segmentation · SIFT · SVM
Blood smear · Microscopic · Feature extraction

1 Introduction

Malaria is an endemic and most serious infectious disease next to tuberculosis throughout the world. Africa, Asia, South America, to some extent in the Middle East and Europe are affected by the disease [1]. Plasmodium species which affect humans are: Malariae, Ovale, Vivax, Falciparum and recently Knowlesi. The only species that is potentially fatal is Plasmodium Falciparum according to Center for Infectious Diseases (CDC) report [2, 4].

F. Mekuria et al. (Eds.): ICT4DA 2017, LNICST 244, pp. 24–33, 2018.
https://doi.org/10.1007/978-3-319-95153-9_3

The distribution of Malaria in Ethiopia can be found in places where the elevation is less than 2300 m above sea level, as can be shown in Fig. 1. The transmission of Malaria is seasonal and hence reaches its peak from September to December following the rainy summer season [12].

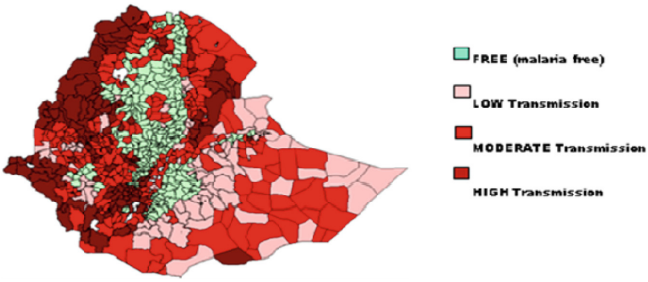


Fig. 1. Map of malaria strata in Ethiopia (©2014) [12].

The two widely known species of Plasmodium in Ethiopia are Falciparum (77%) and Vivax (22%). Relative frequency varies in time and space within a given geographical range. Plasmodium Malariae and Ovale are rare and less than 1%. 60% of the population lives in lowland areas where Malaria can easily spread. Out of the overall population more than 11 million (13%) is under high risk of the infectious disease.

The economic impact in the countries which are affected by Malaria is huge. According to World Health Organization (WHO), total funding for Malaria was estimated to be US$ 2.9 billion in 2015. Governments of endemic countries provided 32% of total funding. According to different studies, 40% of public health drug expenditure is allocated for Malaria, 30% to 50% of inpatient admissions and up to 60% of outpatient health clinic visits are due to Malaria [2, 3], not to mention the humanitarian and non-governmental organizations supporting in different ways.

The reasons for the death toll in the aforementioned regions are due to convenient tropical climate for the growth of the parasites, inadequate technology to combat the disease, illiteracy, and poor socio-economic conditions which make access difficult to health and prevention resources [3]. So, to prevent and eradicate Malaria by the help of technological applications, this paper tries to address image processing techniques and machine learning based identification and classification algorithms which facilitate the diagnosis process.

Mosquito consumes human blood by biting, sporozoites circulate in the blood stream and finally move to the liver where they multiply asexually for some time. In the liver merozoites are regenerated and then invade RBCs [4, 5]. Within RBC the parasite either grows until it reaches a mature form and breaks the cell to release more merozoites into the blood stream to conquer new RBCs or it may grow to reach asexual form named gametocyte and be taken by a mosquito to infect another person where it sexually regenerates to produce sporozoites [6].

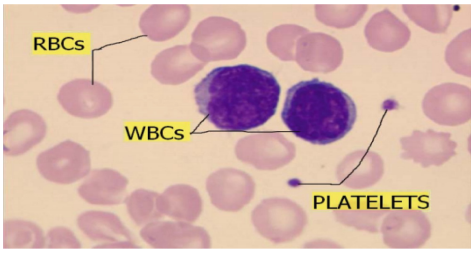


Fig. 2. Healthy thin blood film image with RBCs, WBCs and Platelets [9]. (Color figure online)

Conventionally, Malaria parasite diagnosis is done by visual detection and recognition of the parasite in a Giemsa (the widely used staining technique) stained sample of blood through a microscope. Blood is a combination of Plasma, RBC, White Blood Cells (WBCs), and Platelets [7]. In an infected blood, not only the blood cell components but also the parasites with the different life stages [8] can be detected.

WBCs, Platelets, Plasmodium species and artifacts are deeply stained and appear as dark blue-purplish whereas RBCs are less stained leaving a bright center (patch) with lightly colorized intensity, as shown in Fig. 2. Based on the variation of stain, which in turn tells us the intensity variation, the parasites can be analyzed. However, the quality of the stained image varies according to the available illumination used during acquisition. Malaria can also be diagnosed using Rapid Diagnosis Test (RDT) or Microscope. Microscopic diagnosis is the gold standard which requires special training and considerable expertise. It involves examination of Giemsa stained thick or thin blood film using a light microscope. The method is labor intensive, time consuming and accuracy depends on experience of experts at the field. Hence, automating the process is important to provide an accurate, reliable and objective result [10]. Furthermore, fast diagnostic method is essential for control and eradication of the disease once and for all. Here, an automatic diagnosing of Malaria, which uses image processing and machine learning algorithms has been presented in order to classify and detect the parasite species.

Table 1, depicts comparison of manual, RDT and Computerized diagnosis of Malaria. Using RDT the diagnosis can be performed in about 15–20 min and requires no special training, equipment or electricity. Detection sensitivities of RDTs are comparable to microscopic diagnosis for a larger number of parasite density. Nevertheless, they do not provide quantitative results. In addition, cost of RDT examination is higher than microscopy. On the other hand, computerized diagnosis can provide more consistent and objective results compared to manual microscopy. For instance, the time needed for examination using mobile devices is less than one minute [18], which implies the diagnosis can be done instantly. Generally, automated diagnosis can detect a large number of parasites per microliter, needs no special training and outperforms in both accuracy and computational time than the others.

The rest of the paper is organized as follows, Sect. 2 presents comparison of image segmentation techniques. Section 3 discusses feature extraction and classification using SIFT & SVM, and Sect. 4 addresses conclusion and future work.

Table 1. Comparisons of manual, RDT and computerized microscopy diagnosis requirements and specifications [14, 15].

	Microscopy (manual)	RDT	Computerized (automated)
Requirements	Electricity (optional)	None	Yes
	Special training	Basic training	Basic training
	Staining chemicals	None	Same + computer + camera
Time	~60 min (subjective)	15–20 min	<1 min [18]
Cost	US $ 0.12-0.40	US$ 0.60–2.50	Similar to manual
Specifications			
Detection threshold	500 par/μl	~100 par/μl	~700 par/μl
Detection of all species	Yes	Some brands	Yes
Quantification	Yes	None	Yes
Species identification	Yes	None	Yes
Life-stage identification	Yes	None	Yes

2 Image Analysis

Analysis of images is the use of computer algorithms to extract some useful information [13]. One of the most critical tasks in image analysis is segmentation of images [11]. In this paper, segmentation and classification methods for malaria infected thin blood smear images are discussed. Clinical image processing can broadly be classified into (i) Macroscopic image analysis, and (ii) Microscopic image analysis [13].

Macroscopic analysis of images analyzes images of human organs such as heart, brain, eye, etc. Microscopic analysis of cells from blood, however, helps to understand the nature of cells, and if there is any parasite present, then it can be diagnosed by analyzing the cells [13]. The focus of the paper is microscopic analysis of blood smear images.

Segmentation of images can broadly be classified into deductive and inductive processing. Deductive processing is analyzing and segmenting of images from a higher level to a lower level which is computationally expensive. On the other hand, inductive technique defines object of interest with specific properties, it filters out objects which have unique parameters. Inductive techniques are computationally better than deductive, the details are depicted in Table 2. The reason being all deductive techniques need conversion of images to other image domains, removal of noise and artifacts, morphological processing, segmentation, post processing, feature extraction and classification. In conventional medical image analysis, different procedures are needed to filter

Table 2. Summary of deductive and inductive segmentation techniques

Main categories	Techniques	Advantages	Disadvantages
Deductive segmentation	K-means clustering, genetic algorithm, thresholding, otsu, harris corner detection	From higher level to lower level processing, have good sensitivity and image is processed step by step	Computationally expensive, sensitive to variation in illumination and it is mage specific
Inductive segmentation	Annular ring ratio (ARR) and modified ARR	No preprocessing, locates only stained components, insensitive to image variation, works with all images and provides accurate location of RBC	Computationally fast but accuracy wise a little bit lesser than deductive

out the RBCs from the rest of the image. Many papers on blood film images for Malaria diagnosis use different types of segmentation techniques for extraction of features and classification as shown in Table 2.

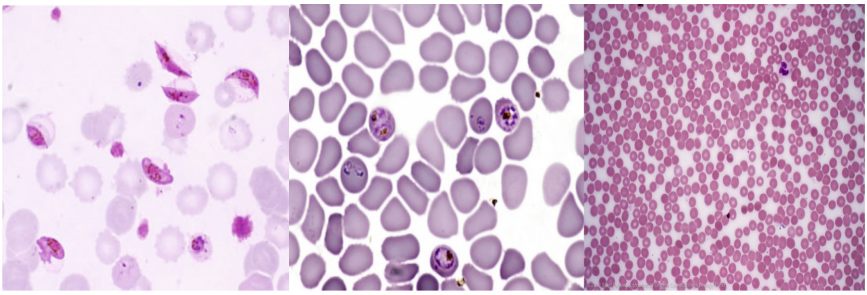
3 Detection of Malaria Parasite with the Help of Machine Learning Algorithms

With the help of Scale Invariant Feature Extraction (SIFT) and Support Vector Machine (SVM) it is possible to detect and classify images with some features into predefined categories or labels.

3.1 Feature Extraction of Images Using Scale Invariant Feature Transformation (SIFT)

This algorithm extracts features and descriptors from all the Gemisa stained images and then clusters using Hough transform. It enables the correct match for a key-point to be selected from a large database of other key-points. The algorithm is invariant to rotation, scale and translation and hence here it is applied to extract Malaria parasite infected RBC images which are deeply stained [16]. The four stages of SIFT have been employed in order to have a well feature extracted image (Fig. 7).

(a) **Scale-space Extrema Detection:-** helps to detect key points from an image by first applying difference of Gaussian at difference scale space and identifying the local minima or maxima of an image as is depicted in Fig. 5(a).
(b) **Key-point Localization:-** following the computation of the difference of Gaussian, each sample point is compared to its neighbor pixels in the current scale space as shown in Fig. 5(b). If the sampled point is maxima or minima then the sampled pixel is labeled as a key-point.



(a) (b) (c)

Fig. 3. Giemsa stained input images: (a) Falciparum, (b) Vivax and (c) Free blood cells.

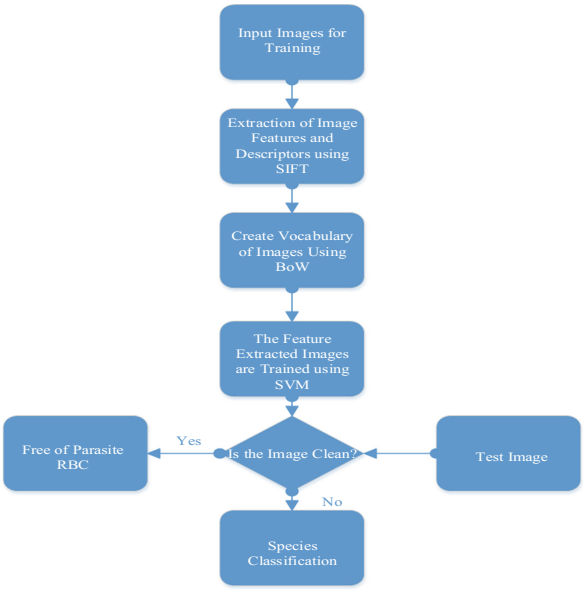


Fig. 4. Flow diagram for diagnosing of malaria using machine learning algorithms.

(c) **Orientation Assignment:-** in order to make features invariant to rotation, orientation is assigned based on local image gradient directions to every feature.

(d) **Key-point Descriptor:-** lastly a descriptor is used at the selected scale (in our case 16 × 16 pixels) in the region around each key-point after an image's location, scale, and orientation to each key-point is known (Fig. 6).

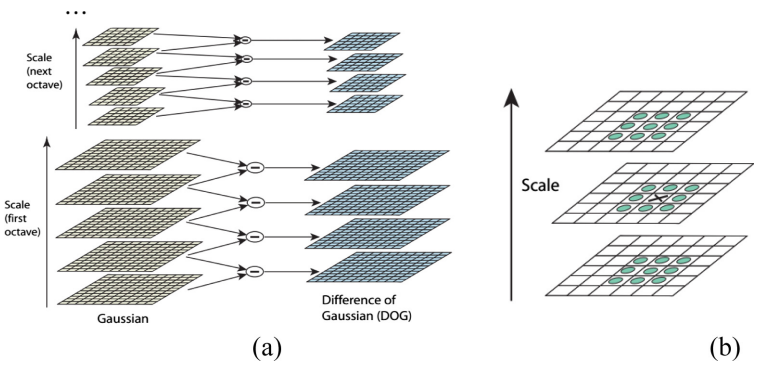


Fig. 5. (a) Repeatedly convolve and down sample by 2 to find difference-of-Gaussian of an image, (b) Computing the maxima or minima of the difference-of-Gaussian from its neighbors [16].

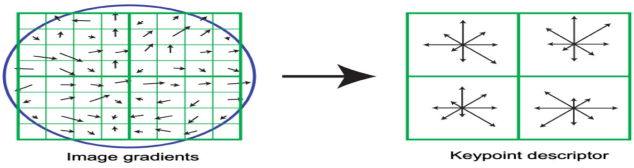


Fig. 6. A 2 × 2 image descriptor array computed from an 8 × 8 set of samples [16].

There are many fast feature detectors and descriptor extraction algorithms like Speed Up Robust Feature (SURF), Robust Independent Elementary Features (BRIEF) and Oriented Fast and Rotated BRIEF (ORB) in image processing but SIFT out performs well when it comes to preserving scale.

The input images are collected from nearby clinics and CDC [6, 19]. The 300 collected images (110 falciparum, 100 vivax and 90 are parasite free images) out of which 90 were used for testing purpose. Hence, each of the species and the parasite free images consists of 30 test images. Classification is done based on type of Plasmodium parasite species. Here, we have considered only parasite free and the two species of the parasite which are prevalent in Ethiopia. However, we are collecting (starting from image acquisition process) more images from the different regions of Ethiopia and we will be studying all the different species using deep learning in our future work. Having our own acquisition process helps us to collect a large number of images and mitigates noises and artifacts which might result due to uncontrolled environment during acquisition from other sources.

By applying SIFT to the input images, the descriptors are shown in Fig. 7(a) and (b). Images with Falciparum (a) and Vivax (b) are labeled with the key-points.

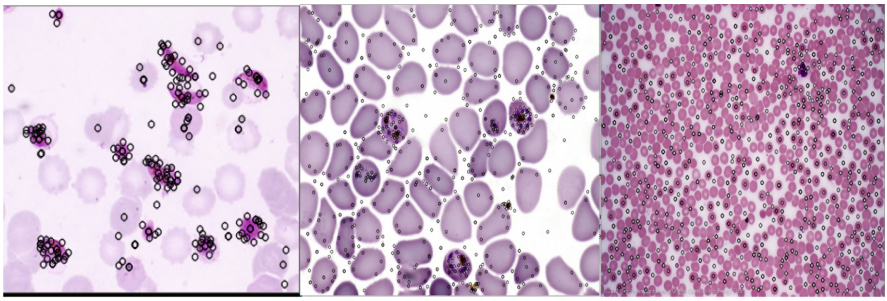


Fig. 7. (a) Image key-points of Falcipurum, (b) Image key-points of Vivax and (c) Free/clean of parasite image

3.2 Creating Bag of Features

Bag of features is an image processing algorithm derived from the well known document classification method known as Bag of Words Model (BoW). Bag of words classifies documents from a large vocabulary of words according to the number of occurrence of words. In a similar fashion, bag of features of images works by creating a vocabulary of image descriptors from the SIFT extracted descriptors.

3.3 Train and Classify Images Using SVM

SVM is a supervised machine learning algorithm that enables to classify new input images based on previously labeled or trained data. It estimates the optimal separating hyper-plane which maximizes the margin of the training data. SVM can use either linear-kernel for a small data set or Gaussian Radial Basic Function (RBF) if the data has large dimension. Since the data set that we have are images, then RBF is used [17].

RBF kernel is very dependent on parameters 'C' (penalty term for misclassification) and gamma, a parameter of a Gaussian that controls the shape. By varying gamma we can increase or decrease the variance. In order to have an optimized values of 'C' and gamma we used opencv's svm.train_auto method which finds the optimal 'C' and gamma values form the given data set. For detecting and classifying Malaria parasite, the features extracted data with SIFT are given to SVM as input for training.

In order to have a well-trained and classified data, Multi-class learning type of classifier is used. Hence, one-against-one method of classification is used to train the collected data into three labels (classes): such as Falciparum, Vivax and parasite free.

After the features are extracted and trained they can automatically classify a test image. If a test image is given to the algorithm, it can be compared with the database of trained images and the image is classified according to the category which is already trained (Falciparum, Vivax or parasite free), that is, the result of the new captured image can be an image that contains Falciparum, Vivax or parasite free. In this work, around 110 images for Falciparum and 100 Vivax, and 90 parasite free images are trained. So far, with a very limited training images, it is reached to a classification accuracy of approximately 78.89% with sensitivity of 80% and specificity of 76.67%. It

can be concluded that as the number of training images are increased (say many hundreds of images are used for training) then the accuracy of the result can be better.

From Table 3, we can see that the sensitivity and specificity values showed acceptable results. However, some images are incorrectly classified, because we don't have enough data for training & testing. Furthermore, the images were not preprocessed, quality of the collected images vary because we have collected them from different sources, [6, 19]. We also noticed that parasite free cells were highly stained and hence they were wrongly classified.

Table 3. Performance measures of the experimentation

Type of component	Falciparum (%)	Vivax (%)	Parasite free (%)	Overall performance of SVM (%)
Sensitivity	83.33	76.67	80	80
Specificity	88.33	85	90	76.67
PPV	78.125	71.875	92.30	82.27
Accuracy	–	–	–	78.89

4 Conclusion

In this paper different conventional image processing techniques are compared in order to detect and classify Giemsa stained microscopic blood smear images of Malaria parasite. Conventional image processing techniques which are studied are preprocessing, filtering, segmenting, feature extraction and classification. We have also implemented SIFT and SVM based classification technique. We have learnt that if there is enough database of images of different species and stages of Malaria parasite, then parasites can be detected and classified with good quality by using machine learning algorithms such as SIFT and SVM.

In our future work we will be collecting a large dataset of Malaria images which include all the five species from the different regions of Ethiopia and employ deep-learning based approach in order to detect and classify the different species and their life stages.

References

1. WHO: Global report on antimalarial efficacy and drug resistance (2000–2010)
2. Korenromp, E., et al.: World malaria report. World Health Organization, Geneva, Technical report (2005)
3. Gallup, J., Sachs, J.: The economic burden of malaria. J. Trop. Med. **64**, 85–96 (2001)
4. National Centers for Disease Control Prevention: Laboratory identification of parasites of public health concern. Division of Parasitic Diseases. Accessed 4 Jan 2017
5. Coatney, G., et al.: The primate malarias. U.S. Department of Health, Education and Welfare (1971)
6. https://www.cdc.gov/malaria/about/biology/

7. Microsoft Corporation: Microsoft encarta encyclopedia (2002)
8. Sherman, I.W.: Malaria: parasite biology, pathogenesis and protection (1998)
9. Kareem, S., et al.: A novel method to count the red blood cells in thin blood films. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1021–1024 (2011)
10. Kareem, S., et al.: Automated malaria parasite detection in thin blood films: a hybrid illumination and color constancy insensitive, morphological approach. Applied DSP and VLSI Research Group, University of Westminster London, United Kingdom (2012)
11. Zhiming, T.: Research on graph theory based image segmentation and its embedded application, pp. 14–24. Dissertation of Shanghai Jiao Tong University, Shanghai (2007)
12. http://www.moh.gov.et/malaria
13. Acharya, T., Ray, A.K.: Image Processing Principles and Applications. Wiley, Hoboken (2005). Arizona State University, Tempe
14. WHO: New perspectives, malaria diagnosis. World Health Organization, Geneva, Technical report (2000)
15. Tek, F.B.: Computerized diagnosis of malaria. Ph.D. thesis, University of Westminster, September 2007
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
17. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
18. Kareem Reni, S.: Automated low-cost malaria detection system in thin blood slide images using mobile phones. Doctoral thesis, University of Westminster, March 2014
19. http://www.mu.edu.et/chs/