ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

Quantification of Parasitemia using Machine Learning Algorithm for Malaria Parasite Detection

Ushasri T[1], Velsundari M[1], Aneeka Amin[1], Nancy Deborah R[2], Alwyn Rajiv S[3]

UG Scholar, Department of Information Technology, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India[1]

Assistant Professor, Department of Information Technology, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India[2]

Assistant Professor, Kamaraj College of Engineering and Technology, Madurai, Tamilnadu, India[3]

ABSTRACT: Malaria is a major cause of death in tropical and sub-tropical countries, killing each year over 1 million people globally. Although effective ways to manage malaria now exist, the number of malaria cases is still increasing, due to several factors. In this emergency, prompt and effective diagnostic methods are essential for the management and control of malaria. Malaria parasites can be identified by examining under the microscope a drop of the patient's blood, spread out as a "blood smear" on a microscope slide. In this proposed work, k-nearest neighbors (KNN) classifier comprising of three features i.e. area, compactness ratio, aspect ratio is used to separate the isolated and compound erythrocytes present in microscopic images of thin blood smear. To develop a systematic method for quantification and classification of erythrocytes in stained thin smear blood films infected with malaria parasite using segmentation operation augmented with KNN algorithm and image processing techniques.

KEYWORDS: Erythrocyte, Thin smear blood films, Microscopic image, Quantification, k-nearest neighbor (kNN)

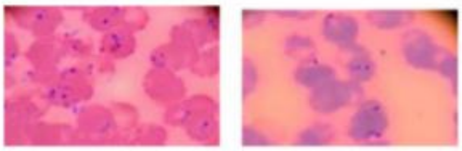


Fig 1. Example of Malaria Parasite in Blood sample

I. INTRODUCTION

During the past few years, various researchers have focused on the development of automatic systems which can analyze the microscopic images of the blood smears. According to the most recent World Health Organization (WHO) report, statistics show that 40% of the world's population is at a risk of contracting Malaria, and more than 240 million infections are encountered every year. Peripheral blood smear examination methods such as Giemsa staining, field staining and Leishman staining provides simple, quantitative, cost-effective (0.03-0.08 USD/test) and species-specific methods for malaria detection. To develop a systematic method for quantification and classification of erythrocytes in stained thin smear blood films infected with malaria parasite using segmentation operation augmented with KNN algorithm and image processing techniques.

II. LITERATURE SURVEY

Object Detection Technique for Malaria Parasite In Thin Blood Smear Images

P.A. Pattanaik, TriptiSwarnkar, DebdootSheet[1] proposes three stage object detection procedure of computer vision with Kernel-based detection and Kalman filtering process to detect malaria parasite. The experiment is conducted on several microscopically preliminary screened benchmark gold standard diagnosis datasets of blood smear images, each 300×300 pixels of Plasmodium falciparum in thin blood smear images. The experimental results on the malaria blood smear image datasets demonstrate the effectiveness of the proposed method over the existing computer vision algorithms. The novelty of the work lies in the application of object detection for malaria parasite identification.

First Polarimetric Investigation of Malaria Mosquitoes as Lidar Target

Detection and Classification of Malaria in ThinBlood Slide Images

Hassan Abdelrhman Mohammed ,Iman Abuel Maaly Abdelrahman[3] works an image processing system that was developed to identify malaria parasites in thin blood smears and to classify them into one of the four different species of malaria. In the first part of the system morphological processing is applied to extract the Red Blood Cells (RBC) from blood images. The second part of the system uses the Normalized Cross-Correlation function to classify the parasite into one of the four species namely, Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale.Accordingly, the RBCs are classified into infected and non-infected cells and the number of RBCs in each image is calculated.

Malaria parasite detection using different machine learning classifier

Adedejiolugboja, zenghuiwang [4] developed a system using stained blood smear images. They employed watershed segmentation technique to acquire plasmodium infected and non-infected erythrocytes and relevant feature was extracted. Six different machine learning techniques for classification are used in the experiments.

Automated Detection of Plasmodium Falciparumfrom Giemsa-Stained Thin Blood Films

Wongsakorn Preedanan, Montri Phothisonotha[5] investigates automated detection of malaria parasites in images of Giemsa-stained thin blood films. The system aim to determine parasitemia based on automatic segmentation, feature extraction and classification methods. Segmentation relies on adaptive thresholding and watershed methods. Statistical features are then computed for each cell and classified using SVM

Samuel Jansson, Peggy Atkinson, Rickard Ignell[2] uses a novel spectral polarimetric optical tomographic imaging goniometer (SPOTIG) to investigate the scattering properties of Anopheles malaria mosquitoes to aid in their detection and identification in entomological lidar applications. In lidar measurements, light scattered by insect wings can be separated from light scattered by insect bodies due to the oscillatory wing beats.

binary classifier. Accuracy of classification is validated based on the leave-one-out cross-validation technique.

III. PROPOSED METHOD

A Python based automated image processing tool for accurate, rapid and user friendly calculation of malarial parasitemia has been proposed. For this purpose, pictures of stained blood samples were collected and analyzed by using proposed automatic tool to calculate parasite infected cells. The created training set was then imported into the Classification Learner application, where the parameters for the model are used and the model was trained and tested which predicts that exact accuracy. The supervised machine learning was used because the training data set contained pre- labeled data and the malarial parasite images. The trained model returned can be imported into the workspace and it would then appear as a user defined function, which can be used for classifying new unknown data.

K-Nearest Neighbors, is one of the simplest machine learning algorithms. KNN is a **non-parametric, lazy** learning algorithm. The k-nearest neighbors algorithm uses a very simple approach to perform classification. When tested with a new example, it looks through the training data and finds the k training examples that are closest to the new example. It then assigns the most common class label to the test example.The normal diagnostic approach to blood disorders is done by blood cell counting and examination of microscopic image of blood smear. In malaria diagnosis, parasitemiaestimation is done which defines the ratio of infected erythrocytes related to total number of erythrocytes in microscopic image.

1. Dataset Collection:

A data set is a collection of data. Data for our project has been taken from Kaggle website which provides thousands of relevant data. Data sets usually come from actual observations obtained by sampling a statistical population, and each row corresponds to the observations on one element of that population. Data sets may further be generated by algorithms for the purpose of testing certain kinds of software. If data is missing or suspicious, we use scikit learn preprocessing which has imputer that will take care of the missing data.

2. Encoding Categorical Data:

We need to encode the data as they are in the text form because it will be complicated for the machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. So we use scikit learn Label Encoder to convert our text data form into numbers. This will help the machines to understand the language and process the data based on the function.

3. Data Preprocessing:

Now the data is split into two categories – Training data and Test data. This process is done to train the machine learning models. The models will be trained by understanding the correlation between the training data. Then the test data are introduced to check the accuracy of the model. The accuracy is given by the prediction of output of the test data. The final step of data preprocessing is to apply the very important feature scaling. It is a method used to standardize the range of independent variables or features of data as every machine learning models use Euclidean distance concept to predict the output.

Euclidean Distance = $\sqrt{((\boldsymbol{x2} - \boldsymbol{x1})^2 + (\boldsymbol{y2} - \boldsymbol{y1})^2)}$

So we transform the data into same standard scale value. Then the data is processed to predict the output.

4. Segmentation:

As a ~~initial step in machine learning~~ process, the blood samples are collected. These blood samples are the microscopic images which are stained using Geisma stain. The samples are provided to predict the output. Now samples are segmented using HSV segmentation. In HSV segmentation, the blood sample will be only in particular colors due to the stain. So we can easily differentiate the infected area from the other. This process will segment the infected area with the help of saturation value.Then the segmented image will be converted into bits because the models will understand the data in terms of numbers.Now the infected area is segmented using image processing technique. It provides the segmented image both in black and white image (0's and 1's) and in original image.

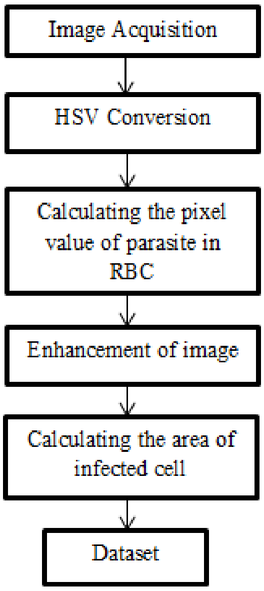
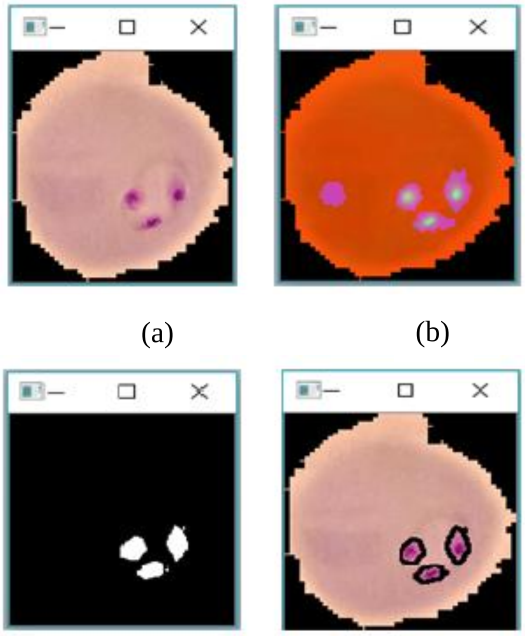


Image Acquisition
HSV Conversion
Calculating the pixel value of parasite in RBC
Enhancement of image
Calculating the area of infected cell
Dataset

Fig 2. Image processing Technique

5. Area Calculation:

Then the area of the infected cell has been calculated with the help of contour function. First the infected area's outer layer is differentiated from the other area to calculate the amount of area of the infected cell. Once the area is calculated it is stored in comma separated file (CSV). This is because every new data that is predicted should become a training data for the next new data. This is performed to provide maximum accuracy of the predicted output and the machine learning model will be well trained for each new data.



(a) (b)

(c) (d)

Fig 3.The reconstructed blood smear images; (a) Original image (b) HSV Segmentation (c) Segmented image (d) Image contoured to calculate infected area.

6. KNN Algorithm:

Now load the data into the machine learning model or algorithm. Initialize K to your chosen number of neighbors. When we decrease the value of K to 1, our predictions become less stable.Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions. Eventually, we begin to witness an increasingnumber of errors. It is at this point we know we have pushed the value of K too far.We usually make K an odd number to have a tiebreaker. Then calculate the distance between the query example and the current example from the data. Add the distance and the index of the example to an ordered collection. Sort the ordered collection of distances and indices from smallest to largest by the distances. Pick the first K entries from the sorted collection. Get the labels of the selected K entries. If regression, return the mean of the K labels. If classification, return the mode of the K labels. This will accurately predict the output of the new data.

Fig 4. Process of KNN Algorithm

We can provide a wide variety of data to the KNN algorithm. It correctly predicts the output and also increases the efficiency of the model. The efficiency increases when each data is processed because once the data has been processed it becomes the training data for the next new data in this model.

IV. RESULTS AND DISCUSSION

The KNN algorithm used here will produce a greater accuracy and lower error rate. The features used in kNN classifier and cell separation process are carried out. The blood cells are converted into HSV format and they are segmented based on the attributes like color, texture etc. The blood cells are separated based on the segmentation process. After the segmentation, the infected area in the blood cell is calculated using the c~~ontour function. As a result t~~he confusion matrix and the accuracy are calculated. The confusion matrix describes the performance of the classification model. The number of correct and incorrect predictions is clearly described.

Two metrics such as accuracy and error rate are used for performance analysis which is given by

Accuracy = (TP +TN) / (TP+TN+FP+FN)

Error Rate = 1 - Accuracy

Where TP = True Positive, FP = False Positive,
FN = False Negative, TN =True Negative.

Once the accuracy and the error rate are calculated we also calculate recall and precision. Recall gives us an idea how often does it predict yes.

Algorithm	**Accuracy**	**Sensitivity**	**Specificity**
SVM	84.2%	89.5%	85%
K-means	90.17%	90.23%	87.3%
kNN	**98%**	**94%**	**92%**

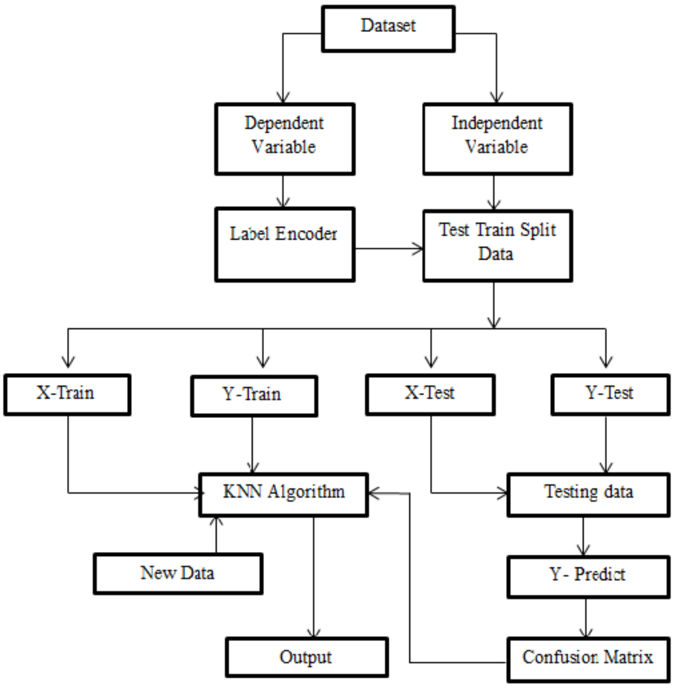
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 3, March 2020

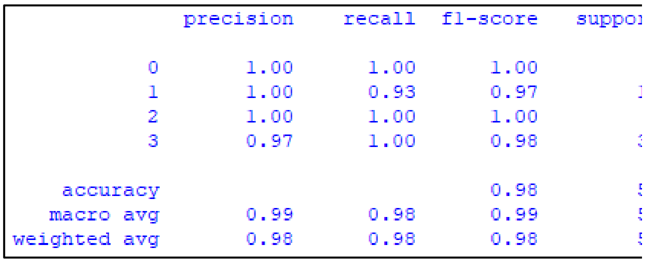


Precision tells us how often it is correct. Then we calculate the f measure score as it clearly explains the effectiveness of the model and its performance.

Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

Fmeasure = (2 * Recall * Precision) / (Recall + Presision)



```
              precision    recall  f1-score   suppo

           0       1.00      1.00      1.00
           1       1.00      0.93      0.97       1
           2       1.00      1.00      1.00
           3       0.97      1.00      0.98       3

    accuracy                           0.98       5
   macro avg       0.99      0.98      0.99       5
weighted avg       0.98      0.98      0.98       5
```

As we analysed we get 0.98 accuracy with high precision and recall. It shows that this method effectively separates the infected blood cells and the normal one.

V. CONCLUSION

During the study, a python (Open CV) based image processingtool was developed and tested for its accuracy and timing. The proposed method isan easy, user friendly, accurate and cost effective method forthe calculation of parasitemia. Erythrocytes separation based on kNN classification comprising of three features has been proposed. The classification process is done by using the three features i.e. area, compactness ratio, aspect ratio. In future, segmentation of the compound cell will be done to determine the number of erythrocyte which is overlapping to each other and furtherimprovements could be done to make the tool more effective.

REFERENCES

[1] P.A. Pattanaik, TriptiSwarnkar, Debdoot Sheet “Object Detection Technique For Malaria Parasite In Thin Blood Smear Images”, IEEE International Conference on Bio-informatics and Bio-medicine(BIBM),2017.

[2] Samuel Jansson, Peggy Atkinson, Rickard Ignell “First Polarimetric Investigation of Malaria Mosquitoes as Lidar Target”, IEEE Journals in Quantum Electronics, 2019.

[3] Hassan Abdelrahman Mohammed, ImanAbuelMaalyAbdelrahman “Detection and Classification of Malaria in Thin Blood Slide Images” International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), 2017.

[4] Adedejiolugboja, Zenghui Wang “Malaria parasite detection using different machine learning classifier” Proceedings of the 2017 International Conference on Machine Learning and Cybernetics, July 2017.

[5]WongsakornPreedanan,MontriPhothisonotha“Automated Detection of Plasmodium Falciparum from Giemsa-Stained Thin Blood Films” IEEE , 2016.

[6] ChowdhurySajadul Islam, Md. SarwarHossainMollah “A Novel Idea of Malaria Identification using Convolution Neural Networks” IEEE EMBS-Conference on Biomedical Engineering and Science(IECBES), 2018.

[7] Mahendra Swain, SandeepDhariwal, Gaurav Kumar “A Python (Open CV) based automatic tool forparasitemiacalculation in peripheral blood smear” International Conference on Intelligent Circuits and Systems, 2018.

[8] BarakaMaiseli, Jiangyuan Mei,HuijunGao, and Shen Yin, Baraka Maiseli“An Automatic and Cost-Effective ParasitemiaIdentification Framework for Low-End MicroscopyImaging Devices”, International Conference on Mechatronics and Control (ICMC), 2014.

[9] CorentinDallet, Saumya Kareem, Izzet Kale“Real Time Blood Image Processing Application forMalaria Diagnosis Using Mobile Phones”, IEEE, 2016

[10] Yuming Fang1, Wei Xiong2, Weisi Lin1, Zhenzhong Chen3“Unsupervised Malaria Parasite Detection Based on Phase Spectrum” IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011.