

Malaria Analysis and Detection Using Machine Learning Approach

¹Prof. Samir Kumar Bandyopadhyay and ²Shawni Dutta

¹GLA University, Mathura, India

²Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India.

***Corresponding Author:**

Prof. Samir Kumar Bandyopadhyay

GLA University, Mathura, India

Type of Publication: Original Research Paper

Conflicts of Interest: Nil

ABSTRACT

This work particularly contributes towards medical image processing. Medical images are generated by several imaging modalities. Such images are interpreted and analysed by some medical professional. Based on the source and imaging modalities the professional may be a radiologist (in case of X-rays, CT, MRI imaging modalities) or a pathologist (microscopic images pertaining to a cell or tissue sample). However, a large number of images are generated that is needed to be analysed with limited number of human resource. Automated computer application or Computer Aided Diagnosis (CAD) systems can provide adequate analysis that can multiply and maximize the limited human resource for correct analysis. Effective, timely and accurate analysis contribute to correct inference that is vital for diagnosis of disease. This paper contributes to this domain of CAD system for Malaria detection from digitized images of PBS (thin smear) slides.

Keywords: Blood Smear; CAD; Medical Imaging; Machine Learning

INTRODUCTION

Malaria disease is common in tropical countries across the globe. Most of the infection is less harmful but some of the infections may result in severe condition and mortality. Malaria is often called “King of Diseases”. Normally this disease takes pandemic proportions during the rainy seasons, where an abundance of stagnant water gives rise to a population of mosquitoes. The disease persists and finds its way to expression and growth among human being.

Economically backward people are most affected by the disease and they have less access to the advancement in preventive medicine and treatment tools. Economists view the interaction of malaria and poverty with the perspective of delineation of economic growth of both individual and nation. Microeconomic analysis of household income and indirect/direct impact of malaria can be up to 10% of annual income of the household. Macroeconomics in

the national level can reduce per capita Gross Domestic Product (GDP) by 50% compared to countries not affected by malaria. The economics were reviewed by economists Sachs and Maloney, who concludes that “where malaria prospers most, human societies have prospered least” [1].

World Health Organization (WHO) manages this disease across the world. They collect data on Malaria and reports annually as ‘World Malaria Report’. The recently published ‘World Malaria Report 2017’, compiles data accumulated in the year 2016 where 216 million cases were reported worldwide (CI 95%) [2]. The number of fatalities reported was estimated to be 4,45,000 deaths globally [2]. Incidence of *P. falciparum* infection was prevalent in Africa whereas *P. vivax* infection was in majority in rest of the world [2]. In India, however, *P. falciparum* infections constitute 66% of the

reported cases while 34% cases are that of *P. vivax* [2].

WHO endorses the view that any case showing malaria like symptoms should be confirmed by laboratory techniques like microscopic examination or Rapid Diagnostic Test kits (RDT) before medication is given [3]. Light microscopy plays an important role for detecting Parasitaemia due to the low cost [3]. Among all the available methods for detection of malaria light microscopy is most predominantly used for Parasitaemia determination [4]. This process was invented during the late 19th century remains the “gold standard” for malaria diagnosis [5-6]. However, use of microscopic diagnosis requires special expertise that is obtained through rigorous training of technicians [6-7].

The smear slides are stained with dye so that it enhances the nucleus region and there is greater contrast for visual identification of White Blood Cell (WBC) and parasite infected Red Blood Cell (RBC). This is, however, a very tedious, time taking process, is subjective and impossible to recreate and there exists “inter and intra-observer variability” in the detection [7-8]. While this requires trained skilled technicians [9], it is a time taking process [10] and accuracy is dependent on the quality of technical expertise and his commitment to the detection process [11], are a major disadvantage of this analysis method. Laboratory diagnosis using microscope achieves 90% accuracy compared with expert microscopy, however, the accuracy level diminishes in the field analysis [10]. Quality of slide preparation, degradation of the slide with time [12] and low agreement rates between experts [13] are an additional disadvantage of microscopic analysis.

There are several alternatives to laboratory diagnosis of Malaria by Peripheral Blood Smear (PBS) with light microscopy. Related methods like Quantitative Buffy Coat Method (QBF), use of Rapid Diagnostic Test (RDT) Kits, Serological tests are available. RDTs are use often for screening purpose. Other expensive molecular methods like Polymerase Chain Reaction (PCR), Loop Mediated Isothermal Amplification (LAMP), Flow Cytometry Method Assay, Automated blood Cell Counters (ACC), DNA Micro Array and Mass Spectrometry (MS) are rarely used. Digital Microscopy has gained popularity over light microscopy. The digital microscopes have

advanced optics, LED lights, high resolution digital camera for recording the observations, firmware and software for managing the data obtained. Whole Slide Scanners have also been utilized for pathological observations. The use of digitized image captured from PBS slides using digital microscopes (Virtual Microscopy) or whole slide scanners have introduced a new concept of ‘Digital Pathology’. The digital images can be archived in the Cloud or transmitted through the network for analysis by expert. This has also provided opportunities for the development of automated image analysis software or Computer Aided Diagnosis (CAD) system.

Computerized Image processing or Digital image processing is often referred to as Computer Vision and found its initial application in the study of satellite imagery captured by spy satellites to track enemy movement. Though initial applications of Computer Vision and Pattern Analysis was in defense related research, but now this has found applications in every walks of life. Digital image processing has wide range of application ranging from industrial manufacturing, Geographical Information System (GIS) and also in Medical imaging. Digital image processing and automated analysis of images have significantly contributed to the betterment of several processes that affect human lives.

The contribution is aimed at developing automated CAD applications for Pathology. Malaria disease identification is an age-old problem that significantly affect a huge population. The work presents an approach for automated analysis of Malaria infected slides.

The complexity of the detection problem involves differentiating normal RBC from infected ones. Some mature parasite forms mimics of the WBC in morphology and hence segregating the parasites from the WBC is vital for Parasitaemia estimation (enumerations of infected cells against the normal cells). This research work is focused on detection of parasites within RBC, determining the Parasitaemia and analysing the morphological forms of infection of two species *P. vivax* and *P. falciparum*, that mostly predominant in India. The Figure 1 below shows the blood cell types and different morphological forms of parasites of interest.

1. Review of Related Research Work

The illumination of images and colour density may vary intra/inter dataset. For execution of an algorithm, it often requires some correction to maintain parity of conditions in the images. Das et al. [14-16] adapted the 'Gray World assumption' for illumination correction. To achieve noise elimination, Median filtering has been adapted by most authors like, Ruberto et al. [17], Ross et al. [18], Anggraini et al [19], Das et al. [16], Rosado et al. [20], Predanan et al [21], Bahendwar et al. [22] and Nugroho et al. [23]. Authors Dave et al. [24] and Savkare et al. [25] have used a combination of Median filtering with Laplacian filter for noise removal along with enhancement of the edge region. Adaptive and local histogram equalisation method is used by Sio et al [26-27], Gaussian filter is employed by Arco et al. al. [28] for noise reduction. The authors Reni et al. [29] performed contrast enhancement on grayscale image by finding optimum weights for R, G and B channels. Automated cell clump removal is vital for accurate enumeration. The red blood cell de-clumping is performed by a rule based binary splitting algorithm as proposed by Kumar et al. [30] is also employed by Sio et al. [26] to de-clump red blood cells for accurate enumeration result. Preedanana et al. [21,31] and Bairagi et al. [32] used Watershed transform method for clump splitting (with Euclidian distance transform).

Segmentation is the key for successful parasite detection. The authors Ruberto et al [17], Tek et al. [14], Das et al. [16], Ahirwar et al. [28], Prasad et al [33] and Dave et al. [24-25] and used Mathematical Morphology for determining the size of red blood cells/segmentation of foreground. The authors Reni et al. [29] utilized a modified 'Morphological Closing' operation for artefact removal and preservation of maximum foreground information. Image binarization using Otsu thresholding for image segmentation was performed by authors Das et al. [16], Ahirwar et al. [28], Anggraini et al. [19]. Mehrjou [34], Rosado et al. [20] and Savkare et al. [25]. Dave et al. [24] have used grayscale histogram with Kurtosis (to determine uni/bimodal histogram) and then performed Otsu thresholding. The authors Bairagi et al. [32] have use Otsu thresholding on RGB and HSV colour channels. Preedanana et al. [21] have adopted adaptive histogram thresholding for segmentation. The authors, Savkare et al [25],

Mehrjou [34], in their research citations have implemented Watershed transform with distance transform for segmentation of red blood cells. Watershed transform method is done by author Das et al [16] [35] and Khan et al. [36]. Authors Damahe et al. [37] used Zack thresholding method on the 'V' or 'value' component of HSV image for segmentation. Authors Purwar et al [38] have used Active Contour model for segmentation of red blood cells.

Some researchers [17] determined the presence of parasite by using two distinct methods. Author Halim et al. [39], performed parasite detection using a Variance based approach and separately a Colour Based Co-occurrence Matrix based matching technique. Authors Tek et al [15] used RGB histogram and probability density function to determine parasite region. Toha & Ngah [40], calculated a threshold value to identify parasite region followed by calculating the Euclidean distance to differentiate between each parasite cluster. Makkapati et al [41], segmented chromatin regions by means of Otsu threshold method using HSV colour model and computed distance of red blood cell region and obtained chromatin regions to differentiate from nucleus of white blood cells. Damahe et al. [37] and Dave et al. [24], converted the image to HSV colour space, while Damahe et al. [18,37,42] performed thresholding on the 'S component histogram', Dave et al. [24] utilized the 'Hue channel' for parasite detection. Fang et al [43], used Quaternion Fourier Transform (QFT) to obtain the amplitude and phase spectrum of image and the inverse Quaternion Fourier Transform to locate parasite region. Elter et al. [44], used green and blue channels for obtaining threshold value, followed by morphological Top-Hat to determine parasite region. The authors Nugroho et al. [23] used K-NN classifier with 'S component' in HSV colour space for segmentation. Khan et al [45] performed clustering on the 'b' component of the image converted to Lab colour space for obtaining parasite region.

Some researchers [15, 22,23] work with 20 classes to identify four species and four stage for each species and normal cases using KNN classifier. Ross et al. [18], used geometric and texture features with Back Propagation Neural Network for classification of parasite infected red blood cell. Different histogram features with Support Vector Machine (SVM) and

multilayer perceptron for classification used by other researchers [31].

Khan et al. [36], used different textural features with Feed-forward Back Propagation Neural Network for parasite identification. Anggraini et al [19], used Multilayer perceptron model for classification. The authors Das et al [16], Ghosh et al [42], used Bayesian and SVM classifier for detecting parasite region. Das et al [35], used texture based features with Multivariate Logistical Regression for identifying parasite in thin smear images.

Authors Savkare et al [21,25, 32] implemented SVM (with colour, texture and shape features) for classification of parasite and normal red blood. Colour and Texture features with SVM classifier is used by Rosado et al. [20,46].

2. Dataset

The glass slide of a thin smear of blood contains a spread of vascular tissue of an individual probably containing Malaria infection. For the purpose of species and life-cycle stage classification 1000x magnification is the standard. Figure 2 shows sample images of Dataset #1 and Dataset #2. **Table 1** shows the image specifications and dataset details.

Dataset #1

The database that was acquired from MaMic [46-47] (which is a publicly available database) pertains to snapshots taken from a whole thin blood smear slide scanned at 100X resolution of *P. falciparum* infection. The *P.vivax* dataset is not available publicly and was provided on request.

Dataset #2

The dataset is acquired from the Pathology Department of SSKM Hospital, Kolkata under the supervision of Dr. D. Ckakraorty. The slides were prepared by the Ronald Ross Malaria Centre within the hospital campus.

Dataset #3

A combined dataset of the images obtained from MaMic [47] and SSKM Hospital, Kolkata.

Methodology

At the very onset, an algorithm was developed using unsupervised technique for identification of red blood cell infection. Thin blood smear images were

acquired from the public MaMic database [46-47]. Images were acquired at 40X magnification and 25watts illumination. Once acquired, image noise was corrected. In total a set of 250 images were used for the study. Of the 250 images acquired, 125 images consisted of blood infected with malarial parasite, while the other images consisted of blood smears taken from individuals not infected with malaria.

Images (of size: 1387 x 932 pixel²) acquired from the MaMic database consisted of salt, pepper noise. To correct salt pepper noise 2D Median filtering with 3 by 3 window was performed. Once noise corrected, the RGB 'JPEG' images were converted to Lab Colour Space image. Based on the a and b components, unsupervised K-means clustering was performed to segment out Red Blood Cells from the Geimsa stained thin blood smear images. The methodology has been duly represented in Figure 3(a) and 3(b) respectively.

To overcome the shortcoming of the proposed methodology, digitized thin blood smears were used to predict the presence of malaria parasite using unsupervised and rule based methods. A dataset was developed from the MaMic database. Cochran's sample size estimation was used to decide on sample size. Clumps are identified based on the third quartile bound of the area distribution of the foreground components. Clumps marked out were de-clumped automatically using 2 phase modified watershed algorithm. Based on the values in $YCbCr$ colour space, the image was recoloured and pixel position matching was performed to detect malaria parasite in Figure. 4.

The segmentation algorithm results in the formation of two clusters, namely, the RBC cluster (consisting of normal RBC cell & (infected RBCs incl. artefacts)) and the WBC cluster (consisting broadly of normal WBC and infection). Based on the infection identification algorithms devised, an intersection of the color value based rule methodology and Lab color based unsupervised clustering algorithm, the infected RBC are identified and thereby segregated from normal RBC cells. Parts of the RBC cell (or pixels) that were marked as infection by only one of the algorithm are treated as probable infection. The intersection parts were marked as infection regions. So in particular, the

RBC cell cluster was further subdivided into 2 sub-clusters, cluster of the normal RBC cells and the infected RBC cells. The probably infected cells were returned to the cluster of normal RBC cells. Based on the annotations provided by a registered medical practitioner the performance of the hybrid algorithm for detection of infected RBC was evaluated

The WBC cluster formed as a part of the aforementioned segmentation process consists of normal WBC cell, infection, artefacts and certain RBC cell outliers that are particularly bigger in size (based on Tukey's Hinges) that RBC and have taken a stain color similar to the WBC cell cluster. Features were calculated for each connected component in the predominantly White Blood cell cluster. The features used to segregate a White Blood Cell from Infection and an outlier Red Blood cell can be divided into 2 broad groups: namely clinical features and features used by other computer science researchers for identification of malaria parasite infection. For either of the two infection clusters, the feature space dimensionality is reduced to a set of 50 features using conditional mutual information maximization algorithm. Now the 50 feature strong dataset of 1410 MaMic database images, 1320 images from the acquired dataset and mixed dataset consisting of 2730 images were all subjected to 10 folds' cross validation to prevent over fitting of the data model developed.

Classification for either infected cluster is performed in three different ways to test the performance among singular classifiers and also compare their performance to ensemble classifiers. Again performance of differently created ensemble classifiers was also compared. The classification strategy that was used has been duly represented in Figure 5.

Results

In accordance with the proposed methodology, the algorithm can particularly be classified into four essential blocks namely, image pre-processing along with ROI extraction, De-clumping of cluster (red blood cell cluster, mixed red blood cell with white blood cell clusters) along with red blood cell After 2-stage de-clumping red blood cells were segregated from white using a threshold value. Tukey's upper hinge was used as a threshold to segregate red from white blood cells, namely, **Table**

Segregation of WBC from other connected components in the cluster under consideration.

So for species and stage classification two images have been used, infected WBC cluster and the infected RBC cluster. For the RBC infected cluster semi-supervised lab color based clustering was performed for pronounced ring identification within the RBC cell. The **Table 3** represents the clinical and other texture features that have been used for species and stage classification of RBC cluster based infection. Features as recorded in **Table 2** have been used for classification of species and stage classification for WBC cluster. The groups in case of White Blood Cell infected cluster are Vivax Schizont and Vivax Gametocyte (Male, Female) [Two class problem]. For the RBC infected cluster the stages and specie groups are Vivax Ring, Vivax Trophozoite, Falciparum Trophozoite, Falciparum Ring, Falciparum Schizont, Falciparum Gametocyte (Male, Female).

enumeration, Malaria parasite detection and parasite stage classification.

For the first section, the foreground identification accuracy achieved by 3-Means Clustering was found to be 92.19 % across all images in the database against the modified Zack's thresholding [48] which recorded only a value of 63.75 % for the dataset at hand. Again, these values were obtained on images for which illumination was not corrected (**Table 4**).

The second integral section can further be divided into two serially dependent blocks, namely, de-clumping and red blood cell enumeration. Modified watershed based methodology was used for automatic de-clumping of red blood cell clusters as also mixed red blood cell-white blood cell clusters. In coherence with the proposed methodology, it was integral to evaluate the performance of different statistical metrics to clearly mark out the clumps (i.e. red blood cell clumps or mixed white and red blood cell clumps) from the individual cells in the digitized thin blood smear image. **Table 5** provides the performance of the different statistical metrics used for differentiating clumps from solitary cell particles (for MaMic database).

6 represents the performance metric for red blood cell segregation from white blood cells. While the first section forms the basis of the algorithm as a whole, the second block is precursor to Parasitaemia

estimation that is worked on in the third block of the research design. The third block is aimed to identify For Parasitaemia estimation the algorithm identifies a white blood cell based on the presence of nucleus (represented in green) and area value along with other texture features as highlighted in **Table 2**, is marked as normal, the other problem area/s present within a red blood cell or as standalone are retained for further investigation.

As per the proposed model, the performance of the hybrid algorithm towards detection of infected RBC at cellular level was recorded (Accuracy 0.9962, Sensitivity 0.9963, and Specificity-0.9949. In case of the WBC Cluster, a 4 –NN classifier was used to separate out white blood cell, infection, artefact (i.e. clustered platelets) and outlier red blood cell. Figure 7 and 8 provide some of the GLCM texture feature values that were found to be statistically significant

Conclusions

The contribution of the work is particularly two fold. In terms of application software, it stands as a tool to assist medical practitioners at effective detection of malaria parasite as also specie and stage classification. As opposed to other toolkits having similar functionality, this particular tool investigates parasites at cellular level which is much preferred by medical practitioners with significantly adding to the computational overhead. In perspective of computer science, it has been a long standing debate with regard to the predictive power of the classifiers. This paper adds on to the vast domain by putting forth a comparative study of single and ensemble classifiers (both inter and intra) using the same set of normalized filter features in perspective of the computer science domain.

As a result of 10 fold cross-validation, the final or best performance that has been achieved by the system is Accuracy of 0.9889, Sensitivity of 0.9949 and Specificity of 0.9892.

However, the model requires further extensive testing before it can be reliable used within the medical domain as an effective aid to the medical practitioner.

References

1. Sachs J, Malaney P (2002) The economic and social burden of malaria. *Nature* 415: 680-685.

presence of malaria parasite in a thin blood smear image.

in distinguishing a White Blood Cell from a Vivax Gametocyte. **Table 7** represents an excerpt of the 88 feature values that were statistically significant towards differentiating a Vivax Gametocyte from a Vivax Schizont in the MaMiic Database.

The evaluation of the algorithm for infection detection within RBC cell and for segregation of WBC from infection, artefact and outlier RBC using a One vs. all strategy for each of the three databases under consideration has been represented in Table 8.

The final feature set of size 50 is selected using Conditional Mutual Information Maximization Algorithm. **Table 9** represents the final average classification accuracy across all infection classes with single and ensemble classifiers respectively.

2. World Health Organization. (2018). World Malaria Report.
3. Reyburn H (2010) New WHO guidelines for the treatment of malaria. *BMJ* 340: 2637-2637
4. Hanscheid T (1999) Diagnosis of malaria: a review of alternatives to conventional microscopy. *Clin Lab Haematol* 2: 235-245.
5. WHO (1991) Basic Malaria Microscopy. Part I. Learner's Guide, World Health Organization.
6. Kettelhut MM, Chiodini PL, Edwards H, Moody A (2003) External quality assessment schemes raise standards: evidence from the UKNEQAS parasitology sub-schemes. *J ClinPathol* 56: 927-932.
7. Payne D (1998) Use and limitations of light microscopy for diagnosing Malaria at the primary health care level. *Bull. W.H.O* 66: 621-626.
8. Kilian AH, Metzger WG, Mutschelknauss EJ, Kabagambe G, (2000) Reliability of malaria microscopy in epidemiological studies: results of quality control. *Trop Med Int Health* 5: 3-8.
9. Oaks SCJ, Mitchell VS, Pearson GW, Carpenter CCJ (eds) (1991) Malaria obstacles

- and opportunities. National Academy Press, Washington, DC, USA.
10. Thung F, Suwardi IS (2011) Blood parasite identification using feature based recognition. Electrical Engineering and Informatics (ICEEI) International Conference.
11. Pammenter MD Techniques for the diagnosis of malaria. *S Afr Med J* 29. 74: 55-57.
12. Ohrt C, Purnomo, Sutamihardja MA, Tang D, Kain KC (2002) Impact of microscopy error on estimates of protective efficacy in malaria-prevention trials. *J Infect Dis* 186: 540-546.
13. Mitiku K, Mengistu G, Gelaw B (2003) The reliability of blood film examination for malaria at the peripheral health unit. *Ethiopian Journal of Health Development* 17: 197-204.
14. Tek, F. B., Dempster, A. G., & Kale, I. (2009). Computer vision for microscopy diagnosis of malaria. *Malaria Journal*, 8(1), 153.
15. Tek, F. B., Dempster, A. G., & Kale, I. (2010). Parasite detection and identification for automated thin blood film malaria diagnosis. *Computer Vision and Image Understanding*, 114(1), 21-32.
16. Das, D. K., Ghosh, M., Pal, M., Maiti, A. K., & Chakraborty, C. (2013). Machine learning approach for automated screening of malaria parasite using light microscopic images. *Micron*, 45, 97-106.
17. Di Ruberto, C., Dempster, A., Khan, S., & Jarra, B. (2002). Analysis of infected blood cell images using morphological operators. *Image and Vision Computing*, 20(2), 133-146.
18. Ross, N. E., Pritchard, C. J., Rubin, D. M., & Dusé, A. G. (2006). Automated image processing method for the diagnosis and classification of malaria on thin blood smears. *Medical & Biological Engineering & Computing*, 44(5), 427-436. doi:10.1007/s11517-006-0044-2
19. Anggraini, D., Nugroho, A. S., Pratama, C., Rozi, I. E., Aulia Arif Iskandar, & Reggio Nurtanio Hartono. (2011). Automated status identification of microscopic images obtained from malaria thin blood smears. *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*.
20. Rosado, L., Costa, J. M., Elias, D., & Cardoso, J. S. (2016). Automated Detection of Malaria Parasites on Thick Blood Smears via Mobile Devices. *Procedia Computer Science*, 90, 138-144.
21. Preedanana, W., Phothisonothai, M., Senavongse, W., & Tantisatirapong, S. (2016). Automated detection of plasmodium falciparum from Giemsa-stained thin blood films. 2016 8th International Conference on Knowledge and Smart Technology (KST).
22. Bahendwar, Y. S., & Chandra, U. K. (2015). Detection of Malaria Parasites through Medical Image Segmentation Using ANN Algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, 5(7), 1063-1067. Retrieved from www.ijarcsse.com
23. Nugroho, H. A., Akbar, S. A., & Murhandarwati, E. E. (2015). Feature extraction and classification for detection malaria parasites in thin blood smear. 2015 2nd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE).
24. Dave, I. R., & Upla, K. P. (2017). Computer aided diagnosis of Malaria disease for thin and thick blood smear microscopic images. 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN).
25. Savkare, S., & Narote, S. (2015). Automated system for malaria parasite identification. 2015 International Conference on Communication, Information & Computing Technology (ICCICT).
26. Sio, S. W., Sun, W., Kumar, S., Bin, W. Z., Tan, S. S., Ong, S. H., ... Tan, K. S. (2007). MalariaCount: An image analysis-based program for the accurate determination of parasitemia. *Journal of Microbiological Methods*, 68(1), 11-18.

27. Arco, J., Górriz, J., Ramírez, J., Álvarez, I., & Puntonet, C. (2015). Digital image analysis for automatic enumeration of malaria parasites using morphological operations. *Expert Systems with Applications*, 42(6), 3041-3047.
28. Ahirwar, N., Pattnaik, S., & Acharya, B. (2012). Advanced Image Analysis Based System for Automatic Detection and Classification of Malarial Parasite in Blood Images. *International Journal of Information Technology and Knowledge Management*, 5(1).
29. Reni, S. K., Kale, I., & Morling, R. (2015). Analysis of thin blood images for automated malaria diagnosis. 2015 E-Health and Bioengineering Conference (EHB).
30. Kumar, S., Ong, S., Ranganath, S., Ong, T., & Chew, F. (2006). A rule-based approach for robust clump splitting. *Pattern Recognition*, 39(6), 1088-1098.
31. Díaz, G., González, F. A., & Romero, E. (2009). A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. *Journal of Biomedical Informatics*, 42(2), 296-307.
32. Bairagi, V. K., & Charpe, K. C. (2016). Comparison of Texture Features Used for Classification of Life Stages of Malaria Parasite. *International Journal of Biomedical Imaging*, 2016, 1-9.
33. Prasad, K., Winter, J., Bhat, U. M., Acharya, R. V., & Prabhu, G. K. (2011). Image Analysis Approach for Development of a Decision Support System for Detection of Malaria Parasites in Thin Blood Smear Images. *Journal of Digital Imaging*, 25(4), 542-549.
34. Mehrjou, A., Abbasian, T., & Izadi, M. (2013). Automatic Malaria Diagnosis system. 2013 First RSI/ISM International Conference on Robotics and Mechatronics (ICRoM).
35. Das, D., Ghosh, M., Chakraborty, C., Maiti, A. K., & Pal, M. (2011). Probabilistic prediction of malaria using morphological and textural information. 2011 International Conference on Image Information Processing.
36. Khan, M. I., Acharya, B., Singh, B. K., & Soni, J. (2011). Content Based Image Retrieval Approaches for Detection of Malarial Parasite in Blood Images. *International Journal of Biometrics and Bioinformatics (IJBB)*, 5(2), 97-110.
37. Damahe, L. B., Krishna, R. K., Janwe, N. J., & Nileshsingh V, T. (2011). Segmentation Based Approach to Detect Parasites and Red blood cells in Blood Cell Images. *International Journal of Computer Science And Applications*, 4(2), 71-81.
38. Purwar, Y., Shah, S. L., Clarke, G., Almugairi, A., & Muehlenbachs, A. (2011). Automated and unsupervised detection of malarial parasites in microscopic images. *Malaria Journal*, 10(1), 364.
39. Halim, S., Bretschneider, T. R., Li, Y., Preiser, P. R., & Kuss, C. (2006). Estimating Malaria Parasitaemia from Blood Smear Images. 2006 9th International Conference on Control, Automation, Robotics and Vision.
40. Toha, S.F., & Ngah, U.K.,(2007). Computer Aided Medical Diagnosis for the Identification of Malaria Parasites. *IEEE - ICSCN 2007*, MIT Campus, Anna University, Chennai, India. 521-522.
41. Makkapati, V.V., & Rao, R. M. (2009). Segmentation of Malaria Parasites in Peripheral Blood Smear Images. *ICASSP 2009*, 1361-1364.
42. Ghosh, P., Bhattacharjee, D., Nasipuri, M., & Basu, D. K. (2011). Medical Aid for Automatic Detection of Malaria. *Communications in Computer and Information Science*, 170-178.
43. Yuming Fang, Wei Xiong, Weisi Lin, & Zhenzhong Chen. (2011). Unsupervised malaria parasite detection based on phase spectrum. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. doi:10.1109/iembs.2011.6091972

44. Elter, M., Hasslmeyer, E., & Zerfass, T. (2011). Detection of malaria parasites in thick blood films. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society.
45. Khan, N. A., Pervaz, H., Latif, A. K., Musharraf, A., & Saniya. (2014). Unsupervised identification of malaria parasites using computer vision. 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE).
46. Annaldas, S., & Shirgan, S. S. (2015). Automatic Diagnosis of Malaria Parasites Using Neural Network and Support Vector Machine. International Journal of Advance Foundation and Research in Computer (IAFRC), 2, 60-66.
47. [MaMic 1 Series Malaria database].
48. Zack, G. W., Rogers, W. E., & Latt, S. A. (1977). Automatic measurement of sister chromatid exchange frequency. Journal of Histochemistry & Cytochemistry, 25(7), 741-753.

Tables and Figures

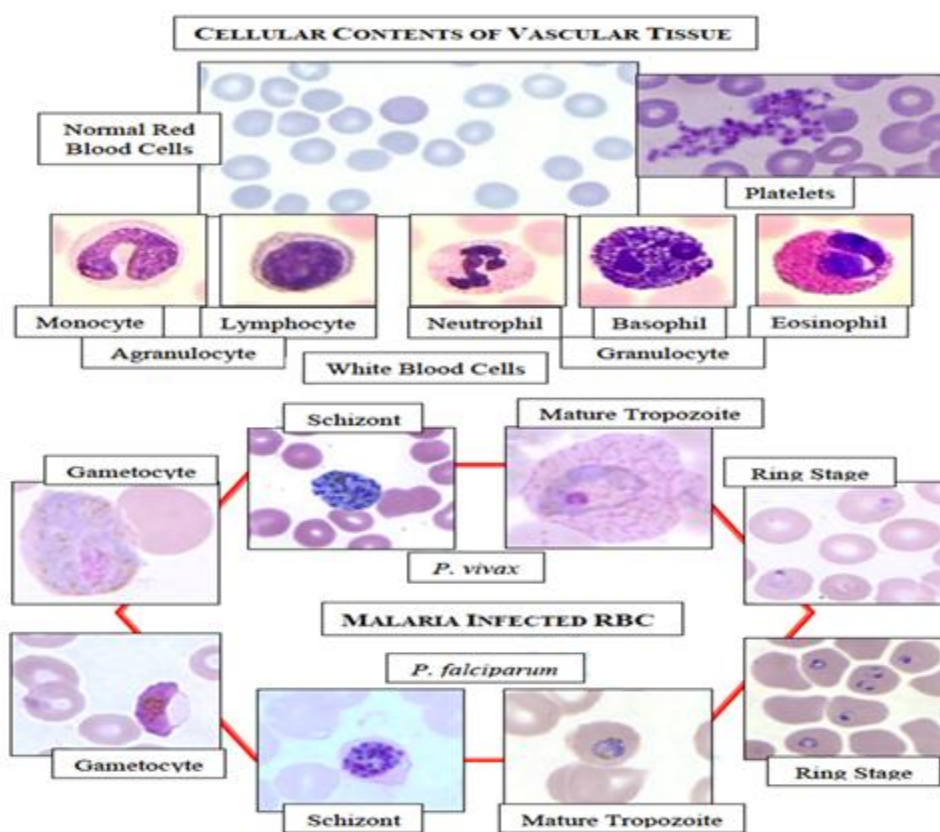


Figure 1: Image showing different cellular components of Blood and different infections that is needed to be identified by the proposed system

Specification	Value
Dataset #1	
Total Number of Scanned Slides	54 slides
Sample size based on Cochran's sample size selection for small datasets	47 slides (23 Normal and 24 infected) 12 <i>P.vivax</i> / 12 <i>P.falciparum</i>
Number of non-overlapping blocks from each slide	30 [Convenience based sampling][From a set of 2790 images] (Understanding Power and Rules of Thumb for Determining Sample Sizes)(each of size 5.08 x 3.39 cm ² [600 x 400 pixel ²])
Image Resolution	300
Magnification used for each Digital Image	100X
Total Number of images used for the Study	30 x 47 = 1410 images
Normal Images	743 (Some slide with infection have normal images due to low Parasitaemia)
Images with Infection	667

P.vivax/P. falciparum	330/337
Vivax Ring – Trophozoite - Schizont - Gametocyte	162 - 115 – 34 – 35
Falciparum Ring – Trophozoite – Schizont – Gametocyte	373 – 38 – 28 – 76

Dataset #2

Total Number of Scanned Slides	33 slides (10 Normal and 23 Infected) 11 P.vivax/ 12 P.falciparum
Number of non-overlapping blocks from each slide	40 (each of size $5.5 \times 2.96 \text{ cm}^2$ [$650 \times 350 \text{ pixel}^2$])
Image Resolution	300
Magnification used for each Digital Image	100X
Total Number of images used for the Study	$40 \times 33 = 1320$ images
Normal Images	400
Images with Infection	920

P.vivax/P. falciparum	440/480
Vivax Ring – Trophozoite - Schizont - Gametocyte	216 - 153 – 46 – 52
Dataset #3	
Total Number of Scanned Slides	80 slides (33-normal, 23-P.vivax, 24-P.falciparum)
Image Resolution	300
Magnification used for each Digital Image	100X
Total Number of images used for the study	1410 + 1320 = 2730 images
Normal Images	743 + 400 = 1143 images
Images with Infection	667 + 920 = 1587 images
P.vivax/P. falciparum	770/817
Vivax Ring – Trophozoite - Schizont - Gametocyte	378 – 268 – 80 – 87

Falciparum Ring – Trophozoite - Schizont 905 – 82 – 67 – 184
– Gametocyte

Table 1: The parameters followed for dataset development with the description of different parasites as observed by experts (ground truth)

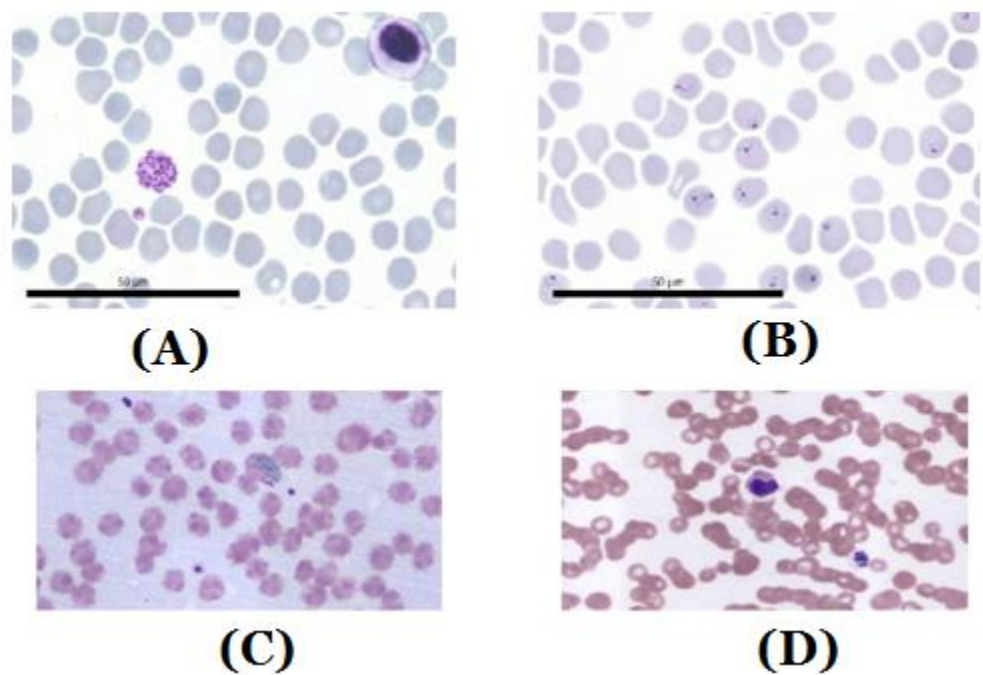
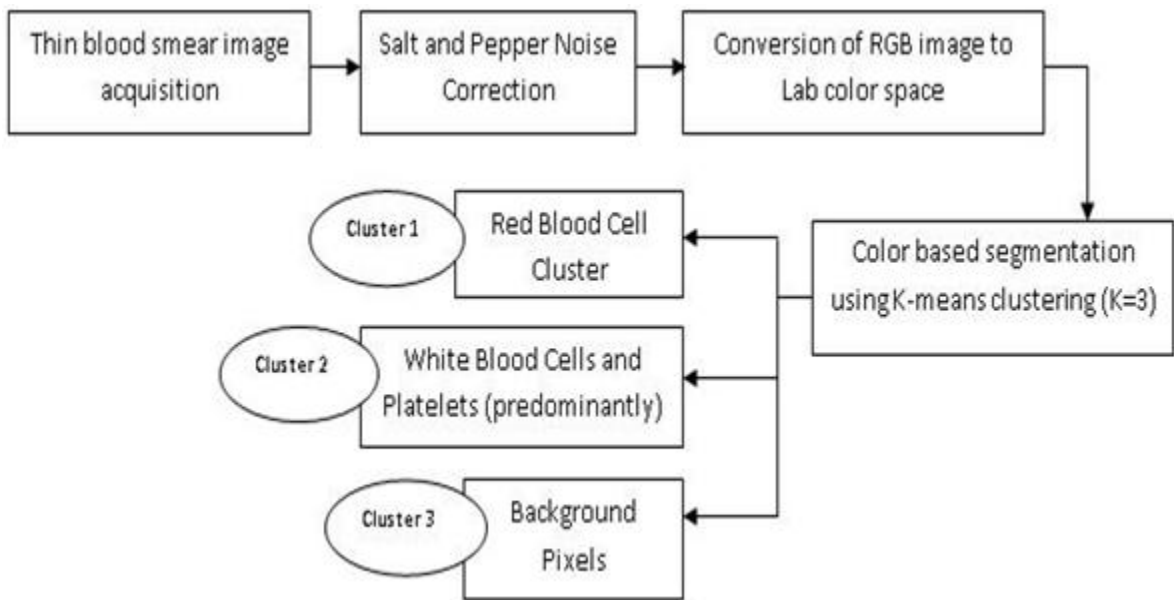


Figure 2 Sample Dataset at 1000X (a) MaMic image showing mature *P. vivax* Schizont, (b) MaMic image showing multiple infection of *P. falciparum* rings, (c) Hospital supplied slide



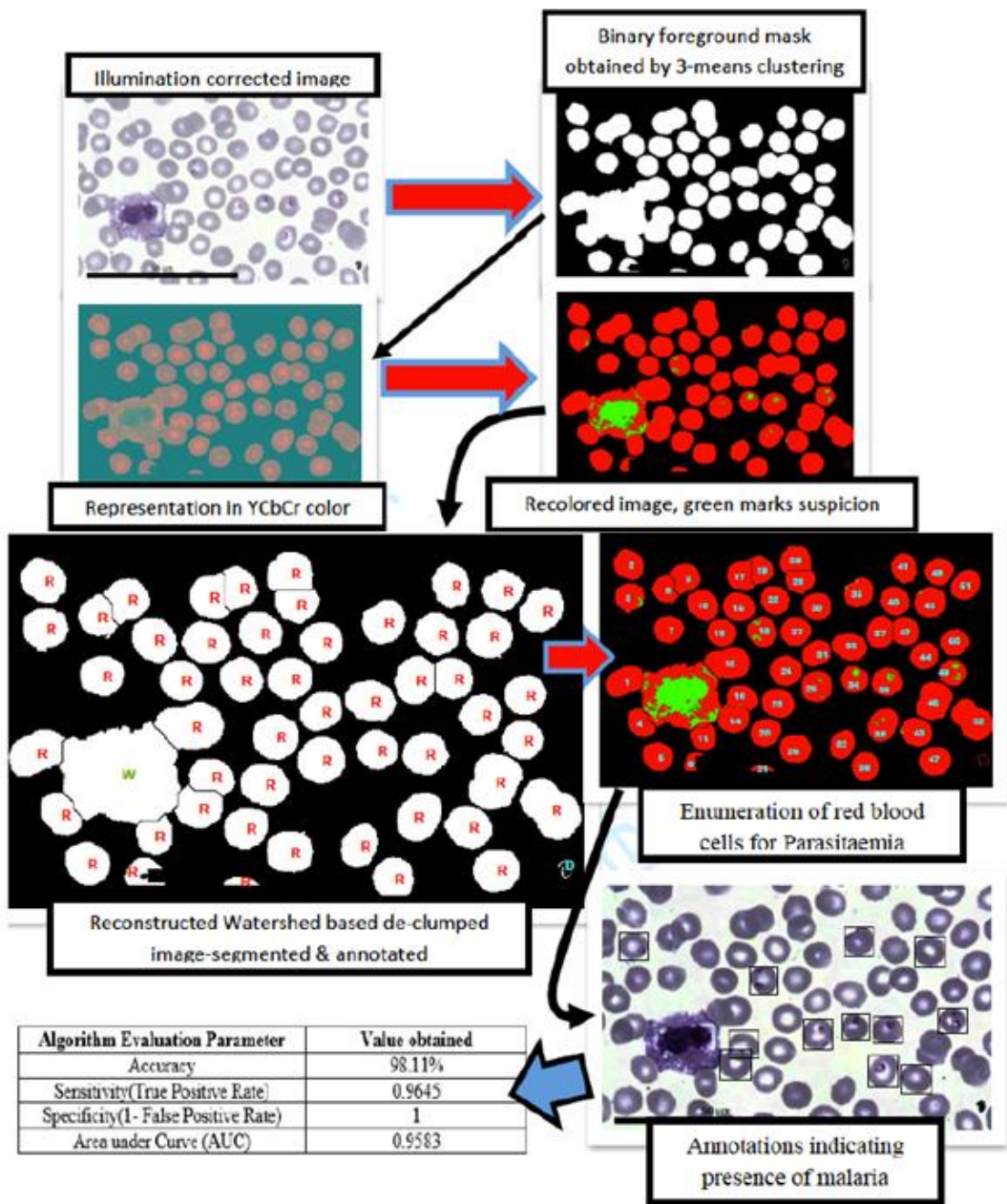


Figure 4: The diagrammatic representation of the parasite detection with results

Feature Name	Number of Features
Nuclear Mass Presence	Number of color coded nuclear mass in a colored component ; Present/absent feature
	(2 features)

Difference of area from Median Area of complete uninfected RBC in a radius of 2 x major axis length of the component area under consideration	Numerical feature, a negative value indicates the size of the area under consideration is smaller than the area of the average uninfected RBC in the given area and				
	likewise for positive value (1 feature)				
Proportion of marked out infection to	Numerical float value(1 feature)				
normal cell area(ratio of color pixel values)					
Features from other authors					
Texture features – Tamura Features	3 features for each grayscale cell component				
Texture features – GLCM	88 features	for	each	Grayscale	Cell
	component				
Texture features – Gabor Features	Not used				
Total Features	94 features				
Clinical Features					
Feature Name	Number of Features				
Nuclear Mass Presence	Number of color coded nuclear mass in a colored component ; Present/absent feature				
	(2 features)				
Difference of area from Median Area of complete uninfected RBC in a radius of 2 x major axis length of the component area under consideration	Numerical feature, a negative value indicates the size of the area under consideration is smaller than the area of the average uninfected RBC in the given area and				
	likewise for positive value (1 feature)				
Proportion of marked out infection to	Numerical float value(1 feature)				
normal cell area(ratio of color pixel values)					
Features from other authors					
Texture features – Tamura Features	3 features for each grayscale cell component				
Texture features – GLCM	88 features	for	each	Grayscale	Cell

	component				
Texture features – Gabor Features	Not used				
Total Features	94 features				

Table 2: Feature List used for segregation of WBC Cluster from other connected components

Features (considering Pathologist perspective)	
Feature Name	Number of Features
Infection Morphology	Area, Perimeter, Eccentricity (3 features)
Cell Morphology	Area, Perimeter, Eccentricity(3 features)
Number of infection instances in a cell	1 feature
Difference of area from Median Area of complete uninfected RBC in a radius of 2 x major axis length of the component area under consideration	Numerical feature, a negative value indicates the size of the area under consideration is smaller than the area of the average uninfected RBC in the given area and likewise for positive value (1 feature)
Proportion of marked out infection to normal cell area(ratio of color pixel values)	Numerical float value(1 feature)
Features from other authors	
Texture features –Tamura Features	3 features for each grayscale cell component
Texture features – GLCM	88 features for each grayscale cell component
Total Features	100 features

Table 3 Feature List used for Specie and Stage Classification from RBC Cluster

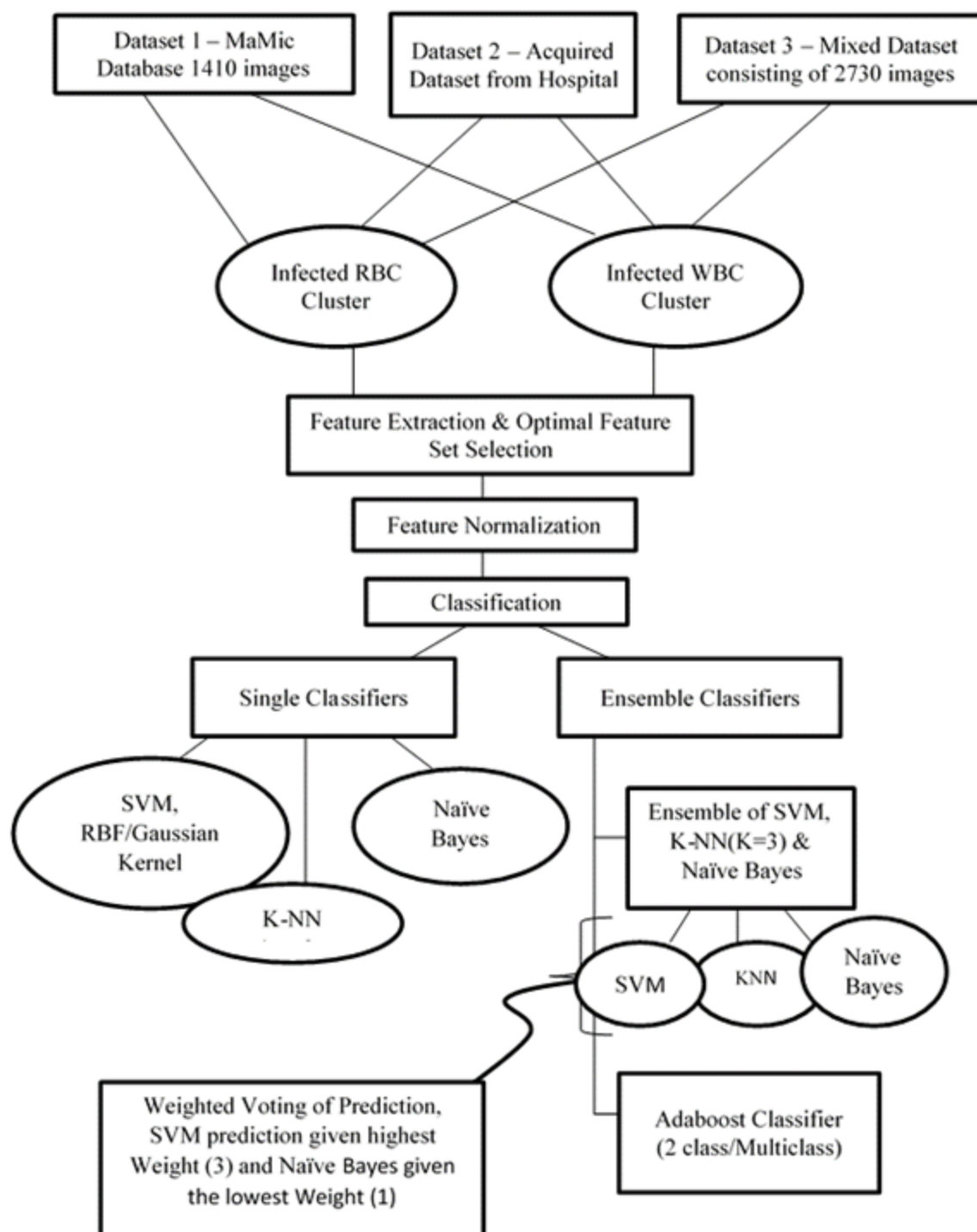


Figure 5 Data-model for Specie and Stage Classification

Method used for Background separation(Without illumination correction of the images)	Accuracy achieved
3-Means clustering	92.19%
Modified Zack's thresholding	63.75%
Algorithm Performance after image illumination correction	
3-Means clustering	98.96%
Modified Zack's thresholding	63.75%

Table 4: Comparative account of the accuracy achieved by the two methods used for image

Statistical Metric	Accuracy	Sensitivity	Specificity
Used/Proposed			
Use of third quartile as a threshold	100 %	1	1
Use of Tukey's upper hinge as a threshold	97.82%	0.3829	0.9982

Table 5: Performance statistics for De-Clump algorithm

Dataset	Accuracy	Sensitivity	Specificity
MaMic Database	0.9925	0.9875	0.9942
Acquired Database	0.9865	0.9944	0.9793

Table 6: Red Blood Cell segregation from White Blood Cell in the digitized thin blood smear for the two datasets under consideration

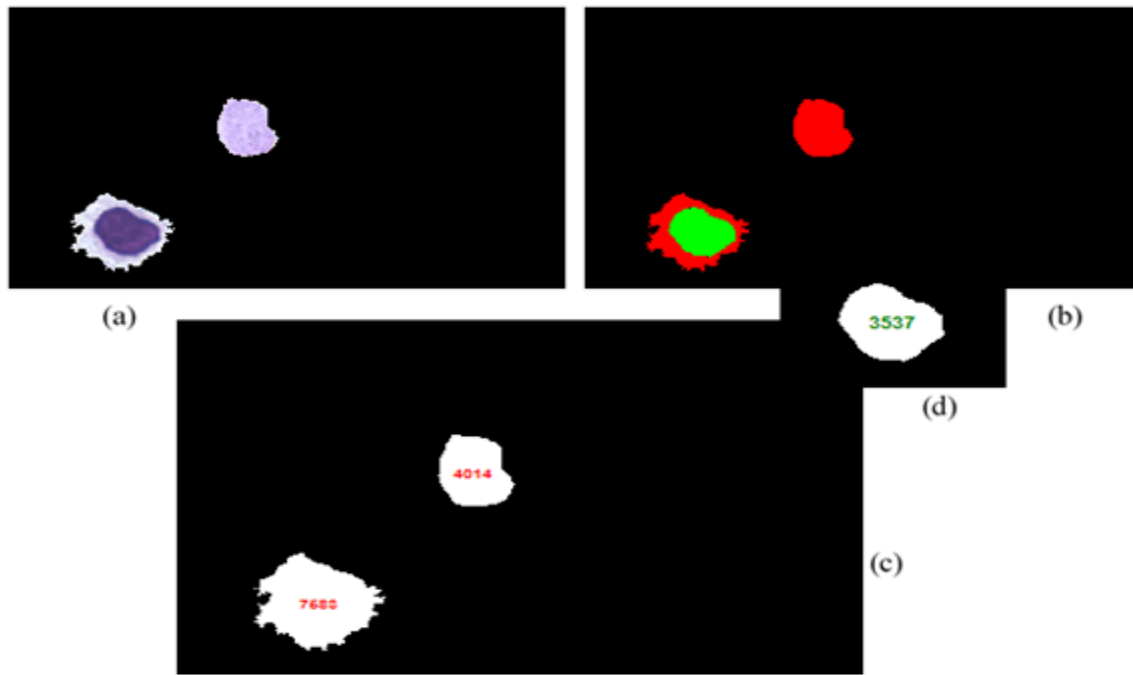


Figure 6 (a) WBC cluster containing a WBC and a vivax gametocyte. (b) Colour coded WBC nucleus and infection, the green represents the WBC nucleus. (c) represents the area of the gametocyte and the WBC, (d) area of the nucleus that has been used as a feature for segregating WBC from Malaria infection in WBC Cluster

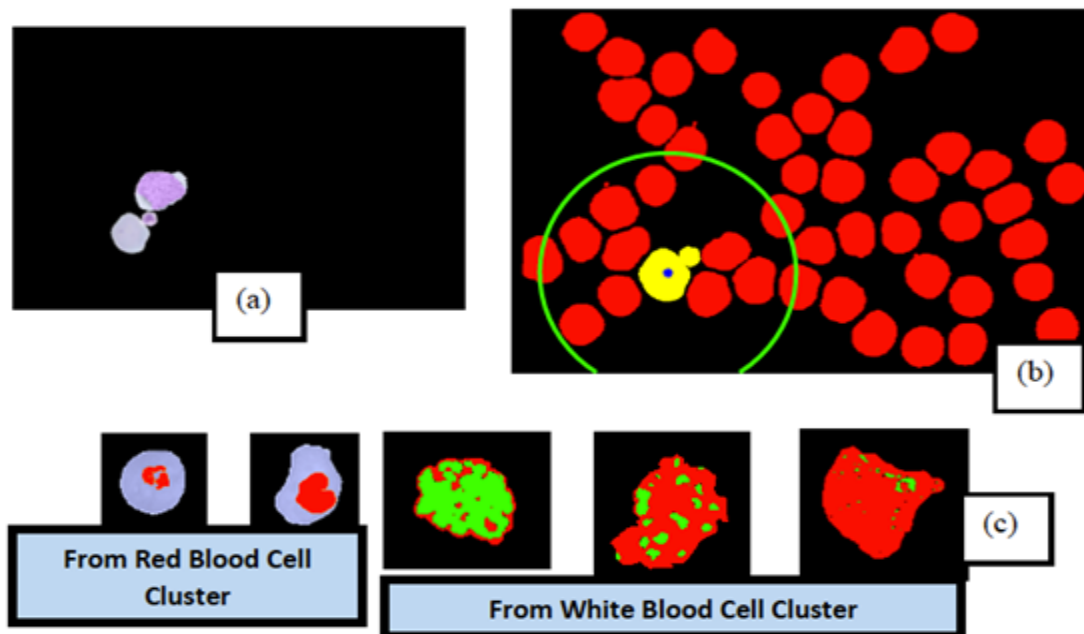


Figure 7: (a) WBC cluster consisting of RBC joined with Platelet Artefact (b) the radius feature used for estimation whether the RBC should be considered as normal or bigger in size (infection/outlier) (c) Different stages of infection P.vivax detected by the proposed

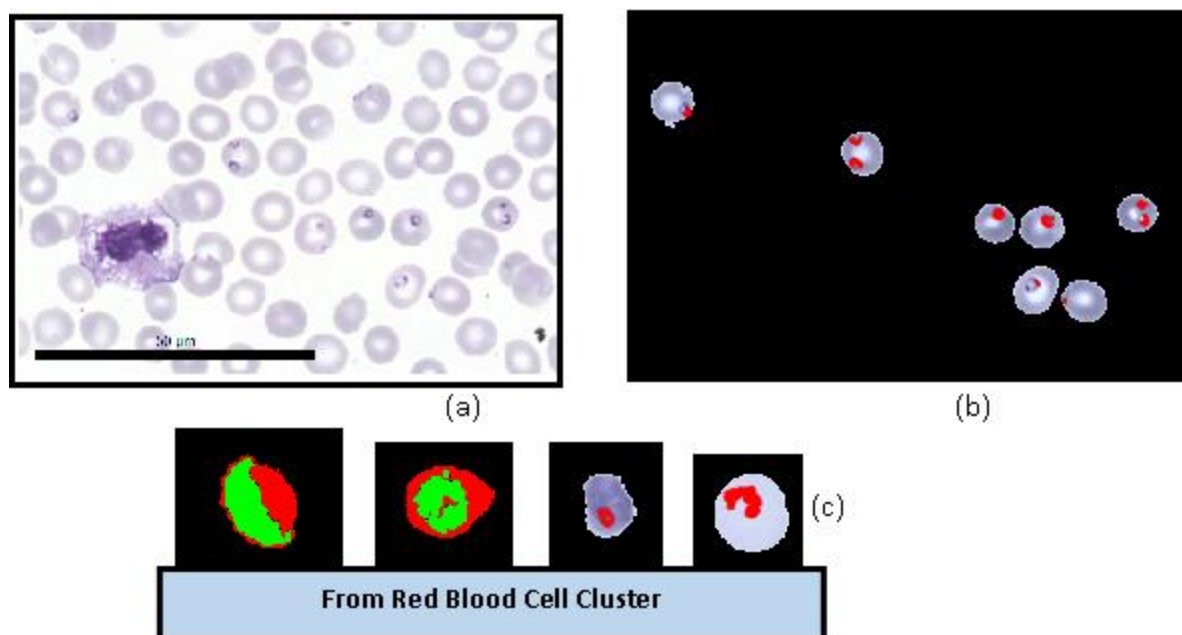


Figure 8: P. falciparum infection at different stages as detected by the proposed algorithm

Features	Prerequisite	Test Performed
Autocorrelation (0°)	Shapiro Wilk Test W(69)=0.892, p=.001<.05	Mann Whitney U Test U=0, p=.000<.05 [significant difference]
Autocorrelation (45°)	W(69)=.887,p=.001<.05	U=0, p=.000<.05 [significant difference]
Autocorrelation (90°)	W(69)=.861,p=.000<.05	U=0, p=.000<.05 [significant difference]
Autocorrelation (135°)	W(69)=.886,p=.001<.05	U=0, p=.000<.05 [significant difference]
Cluster Shade(0°)	W(69)=.735,p=.000<.05	U=0, p=.000<.05 [significant difference]
Cluster Shade(45°)	W(69)=.728,p=.000<.05	U=0, p=.000<.05 [significant difference]
Cluster Shade(90°)	W(69)=.719,p=.000<.05	U=0, p=.000<.05 [significant difference]
Cluster Shade(135°)	W(69)=.731,p=.000<.05	U=0, p=.000<.05 [significant difference]
Dissimilarity(0°)	W(69)=.943,p=.038<.05	U=0, p=.000<.05 [significant difference]
Dissimilarity(45°)	W(69)=.941,p=.030<.05	U=0, p=.000<.05 [significant difference]
Dissimilarity(90°)	W(69)=.896,p=.001<.05	U=0, p=.000<.05 [significant difference]
Dissimilarity(135°)	W(69)=.908,p=.003<.05	U=1, p=.000<.05 [significant difference]
Energy(0°)	W(69)=.864,p=.000<.05	U=0, p=.000<.05 [significant difference]
Energy(45°)	W(69)=.888,p=.001<.05	U=, p=.000<.05 [significant difference]
Energy(90°)	W(69)=.858,p=.000<.05	U=0, p=.000<.05 [significant difference]
Energy(135°)	W(69)=.859,p=.000<.05	U=0, p=.000<.05 [significant difference]
Entropy(0°)	W(69)=.909,p=.003<.05	U=0, p=.000<.05 [significant difference]

Entropy(45°)	W(69)=.923,p=.008<.05	U=11, p=.000<.05 [significant difference]
Entropy(90°)	W(69)=.953,p=.080>.05	Levene's Test(F)=.806(p=.375>0.05) , t(67)=-9.783(p=0.000<0.05) [significant difference]
Entropy(135°)	W(69)=.956,p=.109>.05	Levene's Test(F)=3.521(p=.068>0.05) , t(67)=-9.626(p=0.000<0.05) [significant difference]
Sum of Squares : Variance(0°)	W(69)=.874,p=.000<.05	U=0, p=.000<.05 [significant difference]
Sum of Squares : Variance(45°)	W(69)=.879,p=.000<.05	U=0, p=.000<.05 [significant difference]
Sum of Squares : Variance(90°)	W(69)=.852,p=.000<.05	U=0, p=.000<.05 [significant difference]
Sum of Squares : Variance(135°)	W(69)=.891,p=.001<.05	U=0, p=.000<.05 [significant difference]
Sum Average(0°)	W(69)=.982,p=.741>.05	Levene's Test(F)=.525(p=.473>0.05) , t(67)=7.729(p=0.000<0.05) [significant difference]
Sum Average(45°)	W(69)=.959,p=.139>.05	Levene's Test(F)=.134(p=.716>0.05) , t(67)=9.726(p=0.000<0.05) [significant difference]
Sum Average(90°)	W(69)=.974,p=.434>.05	Levene's Test(F)=2.656(p=.111>0.05) , t(67)=9.001(p=0.000<0.05) [significant difference]
Sum Average(135°)	W(69)=.947,p=.049<.05	U=22, p=.000<.05 [significant difference]
Sum Variance(0°)	W(69)=.835,p=.000<.05	U=0, p=.000<.05 [significant difference]
Sum Variance(45°)	W(69)=.876,p=.000<.05	U=0, p=.000<.05 [significant difference]
Sum Variance(90°)	W(69)=.860,p=.000<.05	U=0, p=.000<.05 [significant difference]

Sum Variance(135°)	W(69)=.831,p=.000<.05	U=0, p=.000<.05 [significant difference]
Sum Entropy(0°)	W(69)=.851,p=.071>.05	Levene's Test(F)=6.861(p=.012<0.05) , t(59.563)=-6.794(p=0.000<0.05) [significant difference]
Sum Entropy(45°)	W(69)=.861,p=.159>.05	Levene's Test(F)=0.355(p=.555>0.05) , t(67)=-6.627(p=0.000<0.05) [significant difference]
Sum Entropy(90°)	W(69)=.917,p=.005<.05	U=16, p=.000<.05 [significant difference]
Sum Entropy(135°)	W(69)=.980,p=.653>.05	Levene's Test(F)=0.070(p=.793>0.05) , t(67)=-6.580(p=0.000<0.05) [significant difference]
Difference Entropy(0°)	W(69)=.899,p=.001<.05	U=0, p=.000<.05 [significant difference]
Difference Entropy(45°)	W(69)=.926,p=.010<.05	U=0, p=.000<.05 [significant difference]
Difference Entropy(90°)	W(69)=.896,p=.001<.05	U=2, p=.000<.05 [significant difference]
Difference Entropy(135°)	W(69)=.923,p=.008<.05	U=0, p=.000<.05 [significant difference]

Table 7: The values of the GLCM features that are significant towards distinction between vivax Schizont vivax Gametocyte from MaMic dataset

Database	Accuracy	Sensitivity	Specificity
RBC Cluster			
MaMic Database	0.9946	0.9963	0.9949
Acquired Database	0.9768	0.9742	0.9773
Mixed Database	0.9860	0.9869	0.9846

WBC Cluster			
MaMic Database	0.9623	0.9609	0.9726
Acquired Database	0.9678	0.9497	0.9705
Mixed Database	0.9649	0.9621	0.9637
Infection in WBC Cluster			
MaMic Database	0.9705	0.9714	0.9710
Acquired Database	0.9565	0.9807	0.9693
Mixed Database	0.9625	0.9770	0.9700

Table 8: Performance statistics for identification of infected RBC from normal RBC Cells and for identification of infected WBC cell cluster using one Vs All Strategy

Classifier(Specie & Stage Classification)	Accuracy [For 50 feature set selected by CMIM]	Sensitivity	Specificity
K-NN(K=3)	0.9653	0.9472	0.9677
SVM(RBF Kernel) C- 1, gamma- ½	0.9759	0.9531	0.9753
Naïve Bayes Classifier (Gaussian distribution assumed)	0.9245	0.9168	0.9344
Ensemble(SVM, 3-NN, Naïve Bayes)	0.9874	0.9925	0.9820
Adaboost (57 trees)	0.9917	0.9948	0.9892

Table 9: Performance statistics for final average classification accuracy across all infection classes with single and ensemble classifiers (at image level)