Detecting Malaria using Machine Learning

T.GRACE SHALINI[1], M.RAJKUMAR[2], C.ARJUN RAJ[3], P.SURYA[4], P.ARPUTHARAJ[5]

Assistant Professor, Dept. of CSE, Velammal College of Engineering and Technology, Madurai, Tamilnadu, India[1]

UG Student, Dept. of CSE, Velammal College of Engineering and Technology, Madurai, Tamilnadu, India[2,3,4,5]

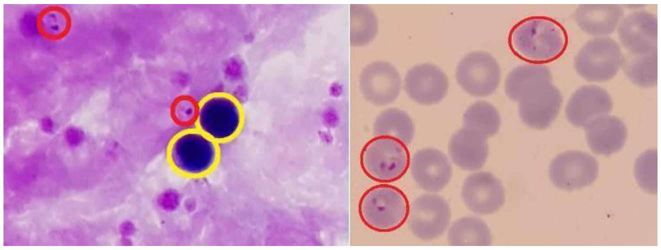
ABSTRACT: This work is used to detect existence of Malarial parasites in BloodCells and to determine how much it is affected. We have developed a machine learning method to do so. Our technique has two processing steps. First, we apply a force based Iterative Global Minimum Screening (IGMS), which plays out a quick screening of a thick smear picture to discover parasite up-and-comers. Then, a customized K-Nearest Neighbor (K-NN) classifies each candidate as either parasite or background. Together with this paper, we make a set of 13780 thick smear images from 500 patients commonly available to research community, Bangalore. This dataset is utilized to prepare and test the AI technique, as portrayed rightnow.

KEYWORDS: Machine learning; K-Nearest Neighbor algorithm; Computer-aided diagnosis; Malaria;

I.INTRODUCTION

MALARIA is one of the globally life-threatening diseases. As stated by World malaria report 2018 [1], roughly 21.9 crore infections were detected in 2017, causing an estimated death of 4,35,000. Diagnosis of the disease manually will be time consuming, also inaccurate. This article gives a review on these methods and talks regarding the present advancements in picture investigation and AI for minute intestinal sickness survey [2], [3]. We compose various methodologies distributed in writing as per procedures utilized to image, picture preprocessing, identification, also cell division, highlight calculation, as well as programmed cell classification [5]. Therefore, developing an automated method is an appealing research objective for improving individualized patient treatment as well as management. It has two big advantages: 1) it can provide a more reliable diagnosis, especially in resource-limited parts, 2) it reduces diagnostic costs. Parasite counts are needed to diagnose, also quantify disease seriousness. In this study, we investigate automatic detection as well as count malarial parasites in digital images of stainsacquired.

A thick blood stain is used to detect the germs. It permits more effective location than a slender blood stain, having around multiple times higher affectability [5]. A meager stain comes about because of laying out a blood drop on a glass slide, and is ordinarily used to separate parasite species and advancement stages. Thick and thin blood stains, as shown in Fig.1, require different processing methods. In thin blood stain, both WBCs and RBCs are clearly visible. A commonplace advance for programmed parasite discovery in slim smears is to initially section RBCs and afterward arrange each portioned RBC as affected or unaffected [5]-[7]. In thick blood smears, however, only WBCs and RBC's nuclei are visible (see Fig. 1(a)). Therefore, parasites need to be detected directly, and a typical step is to first preselect candidates. Then needs to be classified as the germ or background noise. This can be challenging because WBC's nuclei and various non-parasite constituent can absorb stain, creating artifacts that can lead to false results.



(a) Thickbloodsmear (b) Thin bloodsmear

Fig. 1. Examples of thick and thin blood smears. Red circles are parasites and yellow circles are white bloodcells.

II.RELATED WORK

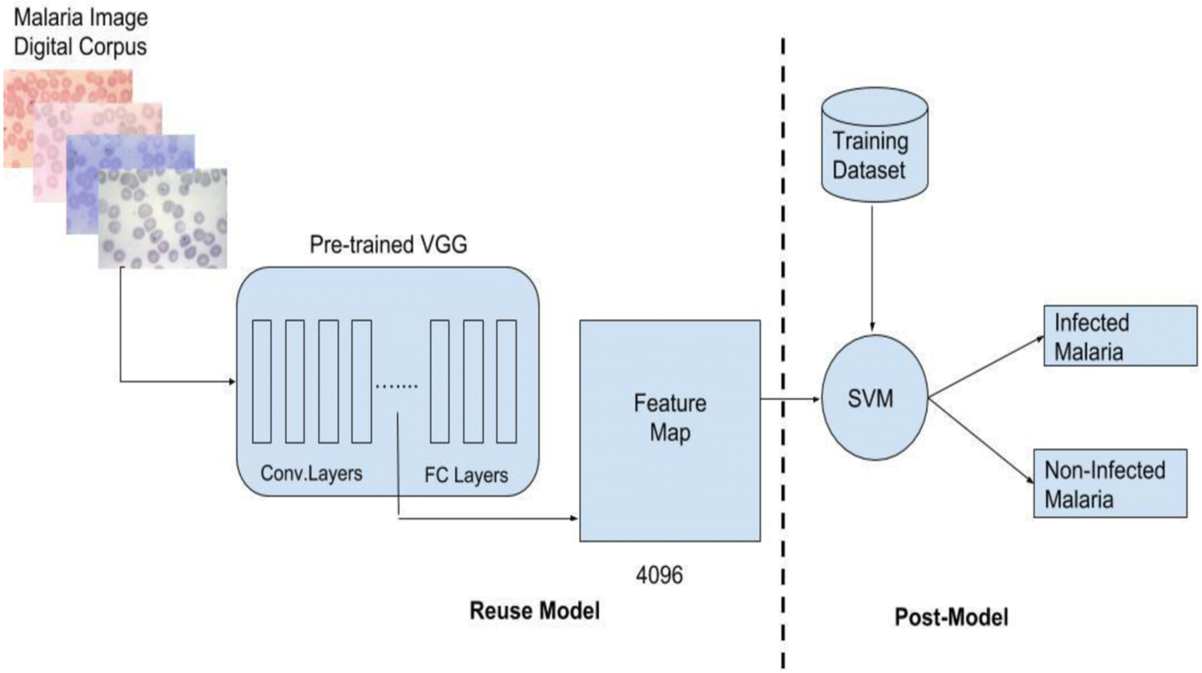
Lately, a few methodologies have been proposed for image handling and examination on blood stains, focusing on computerized recognition. Published literature's reviews may be found in [5],[8],[9]. In the accompanying passage, we give a concise outline of the methodologies for intestinal sickness discovery in thick blood spreads. Traditional perceiving techniques are often performed based on segmentation [10] using thresholding and morphological operations. Kaewkamnerd et al. propose a method using an adaptive threshold on the V-value histogram of the HSV image to extract parasite candidates and WBCs from the background, and then distinguish parasites from WBCs according to their size. Evaluation on 20 images shows that the proposed method achieves an accuracy of 60%. Hanif et al. use an intensity-stretching method to enhance the contrast of 255 thick blood smears, and then use an empirical threshold for segmenting malarial parasites. The authors show qualitative results on different images, in which different empirical thresholds are applied to obtain satisfying segmentation results. Chakrabortya et al. consolidate morphological division with shading data for recognizing parasites in thick blood spreads. Analyses are performed on 75 pictures and fix level assessment shows a fruitful discovery pace of 95% with a bogus positive proportion of 10%. Dave et al. perform histogram-based versatile thresholding and morphological procedure on denoised pictures to recognize RBCs tainted by jungle fever parasites in thin and thick blood spreads.

Patch level evaluation on 87 images shows that the method detects 533 parasites compared to 484 parasites annotated as ground truth. Traditional approaches for parasite detection are simple and fast, whereas they are difficult to extend to large datasets. This is due to the fact that traditional approaches are very sensitive to image variations and that parameters are very often determined empirically. Performance evaluation onpatch level on small datasets (from 20 to 300 images) can change greatly when evaluating on big datasets, on image level or patient level.

Feature-based approaches involve feature extraction and classification based on machine learning techniques . Elter et al. extract 174 features from pre- detected plasmodia candidates and apply a Support Vector Machine (SVM) classifier to the feature set for parasite identification. The creators report an affectability of 97% for 256 pictures on fix level. Purnama et al. separate highlights from histograms of RGB channel, H channel from HSV space, and H channel from HIS space, and afterward utilize Genetic Programming to distinguish parasite type and stage. Their order model on 180 patches accomplishes a normal exactness of 95.58% for parasite distinguishing proof and 95.49% for non-parasite ID. Yunda et al. remove shading highlights, co-event surface highlights, and wavelet-based surface highlights from the pre- sectioned picture, and afterward use Principal Component Analysis (PCA) to lessen excess highlights, trailed by a neural system model for the last grouping.Assessment on 110 pictures shows that the affectability for parasite location is 76.45%. Quinn et al. propose to initially part each picture into 475 arbitrarily covering patches utilizing downsampling and sliding window screening, at that point remove associated segment and minute highlights from the patches, lastly utilize a randomized tree classifier for the arrangement The strategy is assessed on 2903 pictures from 133 patients and produces an accuracy of 90% at a review of 20% on fix level. Rosado et al. utilize a versatile thresholding approach for the parasite location and afterward apply geometry, shading and surface highlights in mix with a RBF bit based SVM classifier for WBC and parasite distinguishing proof. Evaluation on 94 images from 6 patients shows their automatic prediction of parasites has achieved a patch level accuracy of 91.8% along with a sensitivity of 80.5% and a specificity of 93.5%, while their WBC detection achieves 98.2% sensitivity and 72.1% specificity. The feature-based approaches evaluate their performance on patch level. That is, the input sample is a single patch image and the evaluation is typically a patch classification accuracy. However, the ultimate goal for malaria patient diagnosis is to detect and classify all patches (both parasites and false positives)for a patient. A satisfying patch level classification performance does not necessarily assure good performance on image level or patient level.

III.PROPOSED SYSTEM

In architecture can include of structure parts and the sub-structure envolved, that will effort cooperatively to implement the overall structure.There have been works to formalize languages to describe structure architecture, collectively these are called architecture description languages(ADLs).



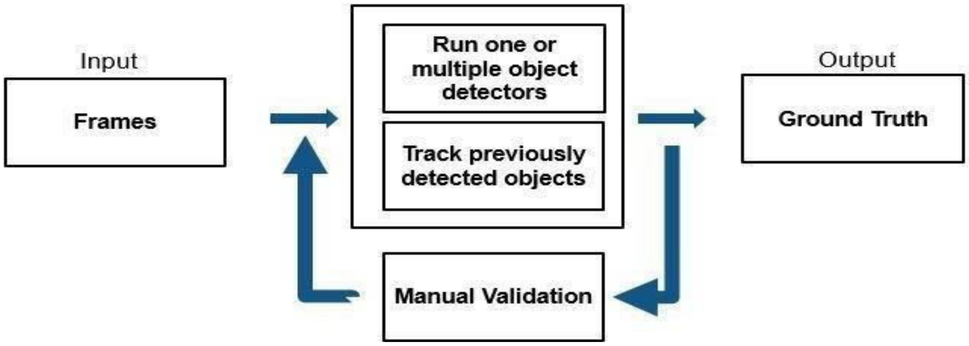
LAYER DESCRIPTION

A.COLLECTION OF PATIENT MEDICAL RECORDS

The data in most of the hospitals are confidential in India. In this project tried to collect data for research purposes but were denied. The data allowed to collect mostly contained name, age and sex which are not enough for prediction. Malaria diseases depends on a lot of variables like hypertension, diabetes, exercise schedules, chest pain etc. Collection of these data was against the hospital rules. In this work therefore had to analyze the data from a dataset repository which had dataset from four different hospitals from four different states. This increases the diversity in the data. Data integration merges data from multiple sources into a coherent data store, like a data warehouse or a data cube. Careful integration of the data from multiple sources helps in reducing and avoiding redundancies and inconsistencies in the resulting data set. This helps in improving the accuracy and speed of the subsequent classification process. Data reduction can reduce the data size by aggregating and eliminating redundant features. Using the data classification techniques, the focus is on specific fields that allow exploration of the data, by selecting and filtering some fields as input, output fields and predictive fields.

B.PRE-PROCESSING OF RECORDS

Our data like most of the datasets contains noise. Since our data was collected from four different databases there were missing data in many cases. This noise and missing data leads to over fitting. Although the data was finally collected, it has a lot of missing data. The data was filled by using median method. The data set was also sparse which made prediction difficult if the scale factor was set to get a refine and more accurate prediction. Cleaning and filtering of the data have necessarily to be carried out with respect to the data in data classification algorithm to avoid the creation of deceptive or inappropriate rules or patterns. To make the data appropriate for the classification process, it needs to be transformed.

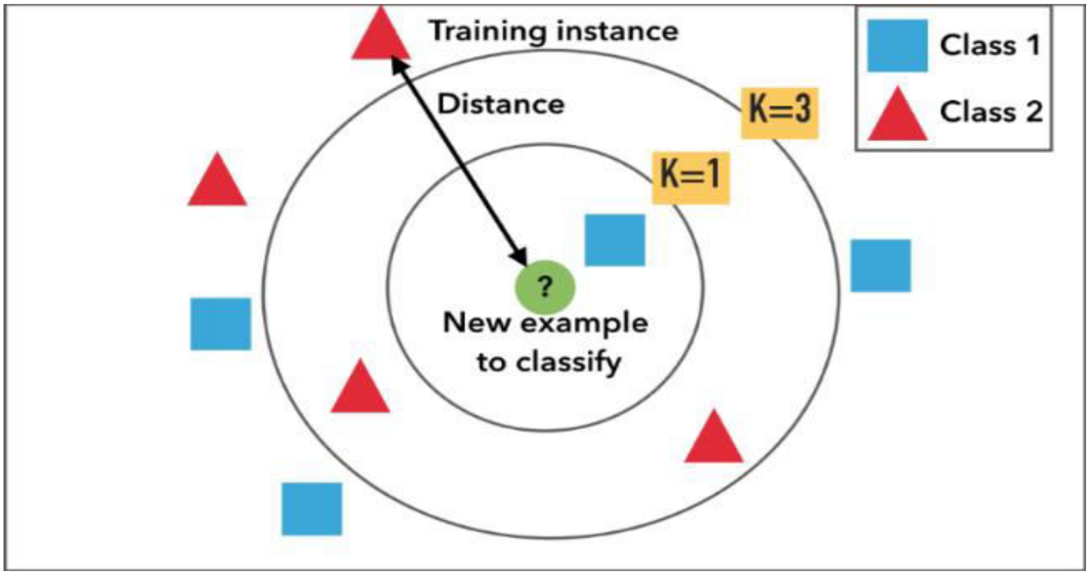

Input
Frames
Run one or multiple object detectors
Track previously detected objects
Output
Ground Truth
Manual Validation

C.PARASITE PRESELECTION USING ITERATIVE GLOBAL MINIMUM SCREENING (IGMS)

IGMS generates RGB parasite candidates by localizing the minimum intensity values in a grayscale image . If only one pixel is localized, a circular region centered at this pixel location with a pre-defined radius of 22 pixels (average parasite radius) is cropped from the original RGB image and is selected as a parasite candidate .If more than one pixel is localized, a new parasite candidate centered at the i^{th} pixel is added when all the distances between the i^{th} pixel and previously selected pixels are larger than 22. Once a parasite candidate is selected, the intensity values inside this region of the grayscale image will be replace by zeros to guarantee the convergence of the IGMS method. The screening stage stops when the number of parasite candidates reaches a given number. Experiments on our dataset of 150 patients show that we can achieve a sensitivity above 97% on patch level, image level, and patient level when using this number. Each parasite candidate is a 44×44×3 RGB patch image, with pixels having a distance greater than 22 to the center set to zero.

D.K-NEAREST NEIGHBOR ALGORITHM

Once the parasite candidates are extracted, we use a K-NN model to classify them either as true parasites or background. In this work, we customize a K-NN model consisting of seven convolutional layers, three max-pooling layers, three fully connected layers, and a softmax layer as shown in batch normalization layer is used after every



ry

K-Nearest Neighbor Algorithm

convolution layer to allow a higher learning rate and to be less sensitive to the initialization parameters , followed by a rectified linear unit (ReLU) as the activation function. Max-pooling layers are introduced after every two successive convolutional layers to select feature subsets. The last convolutional feature map is connected to three fully connected layers with 512, 50, and 2 hidden units, respectively. Between the three fully connected layers, two dropout layers with a dropout ratio of 0.5 are applied to reduce model overfitting. The network is derived from VGG19 by selecting the first six convolutional layers and three corresponding max-pooling layers from the VGG19 architecture to stop the feature maps, followed directly by the fully connected and dropout layers. This shorter network structure provides similar performance while being faster and requiring less memory, which is important for smartphone applications [9]. The output of the K-NN model is a score vector, which gives the probabilities of the input image patch being either a parasite or background. We can obtain a larger or smaller number of predicted parasites by applying an adaptive probability threshold to the score vector. Compared with pre-trained networks such as VGG , GoogLeNet , ResNet-50, our customized K-NN model has several advantages: 1) runtime is reduced by using a smaller set of customizable parameters, with the input size of the model being determined by the average parasite size in thick smear images (44×44×3), which is much smaller than the input size used by the other networks (224×224×3); 2) our smaller network structure with fewer layers is more suitable for smartphones. Since the input size is smaller, our network should in fact be less machine to avoid feature maps that are too small. A smaller network structure with less parameters also avoids over- training on the smaller input space. Compared to the pre-trained networks mentioned above, our customized K-NN model achieves a better accuracy, despite having less network layers, and a shorter runtime. For an input image of 4032×3024×3 pixels, our system can complete the parasite detection within ten seconds (about eight seconds for candidate screening and two seconds for classification) on a standard Android smartphone. Both the smaller set of parameters and the smaller network structure contribute to the reduced runtime.

IV. PARAMETER SETTINGS

A.ACCURACY

Accuracy is the instrument to calculate the accurate value and it is a proportion of accurate state consideration to the total considerations.

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

B.PRECISION

Precision is the magnitude of accurate state positive
considerations to the total state positive consideration and it is a calculate of consistency and reproducibility.

Precision = TP/TP+FP

C.RECALL (SENSITIVITY)

Recall is the proportion of number of relevant instances to the total number of actual relevant instances.

Recall = TP/TP+FN

D.F1 SCORE

F1 Score is the mass mean of Precision and Recall. It gives a better calculate of the wrongly organized cases than the Accuracy metric. Accuracy is used when the accurate positives and accurate negatives are more important F1 Score is used when the inaccurate negatives and inaccurate positives are vital.

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

- **TP-True Positives**
- **FP- False Positives**
- **FN-False Negatives**
- **TN-True Negatives**

V. COMPARISON WITH SVM ALGORITHM

It has been observed that K-NN provides greater accuracy rate. Algorithms like Gradient Descent Ensemble performs good with an accuracy of 79% when compared with SVM which performs only 54%.Convolution Neural Network is non-linear classifier. Support Vector Machine is a linear classifier. K-NN performs good with Visual image recognition where as SVM is used most probably for classification problems. K-NN is a feed forward neural network which is generally used to analyse visual images by processing data with grid like topology. Support Vector Machine is an algorithm that analyse data used for classification and regression analysis.SVM is a supervised learning method that looks at data and sorts it into one of the two categories.

VI.CONCLUSION

In this paper, we are implementing a smartphone based application to detect the malarial parasites in thick blood stain images. Our automated processing pipeline has two stages: parasite screening and classification. An intensity-based Iterative Global Minimum Screening (IGMS) initially conducts a quick screening of entire thick blood stain images to produce parasite candidates. A modified K-NN model then classifies each candidate as a parasite or background. Our experimental results show the practicality of our automated system. Our paper is, to the best of our knowledge, the second paper to develop a smartphone based system for thick blood smear screening, and the first to apply machine learning techniques to detect the parasites in thick smears on smartphones, with patient-level assessment. We make our dataset of 1819 images from 150 patients publicly accessible, as a service to the research community, which will alleviate the problem of lacking training data for automated malaria diagnosis in thick blood smears. Our future research is to enhance the functioning of our system using network ensemble techniques and to reduce its runtime onsmartphones.

REFERENCES

[1]. WHO, World Malaria Report 2018. 2018.
[2]. WHO, Guidelines For The Treatment of Malaria, Third edition. World Health Organization, 2015.
[3]. K. S. Makhija, S. Maloney, and R. Norton, “The utility of serial blood film testing for the diagnosis of malaria,” Pathology, vol. 47, no. 1, pp. 68–70, 2015.
[4]. WHO, Malaria micropscopy quality assurance manual, Version 2. World Health Organization, 2016.
[5]. M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, “Image analysis and machine learning for detecting malaria,” Transl. Res., vol. 194, pp. 36– 55, Apr. 2018.
[6]. Z. Liang, A. Powell, I. Ersoy, M. Poostchi, K. Silamut, K. Palaniappan, P. Guo, M. A. Hossain, A. Sameer, R. J. Maude, J. X. Huang, S. Jaeger, and G. Thoma, “CNN-based image analysis for malaria diagnosis,” in Proc. BIBM, ShenZhen, China, 2017, pp. 493–496.
[7]. S. Rajaraman K. Silamut; M. A. Hossain, I. Ersoy,
R. J. Maude, S. Jaeger, G. R. Thoma, and S. K. Antani, “Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images,” J. Med. Imaging, vol. 5, no. 3, p. 034501, July 2018.
[8]. L. Rosado, J. M. Correia da Costa, D. Elias, and J.
S. Cardoso, “A Review of Automatic Malaria Parasites Detection and Segmentation in Microscopic Images,” Anti-Infective Agents, vol. 14, no. 1, pp. 11–22, Mar. 2016.

[9]. P. A. Pattanaik and T. Swarnkar, “Comparative analysis of morphological techniques for malaria detection,” Int. J. Healthc. Inf. Syst. Informatics, vol. 13, no. 4, pp. 49-65, Oct. 2018.

[10]. S. Kaewkamnerd, A. Intarapanich, M. Pannarat, S.Chaotheing, C.Uthaipibull, and S. Tongsima, “Detection and classification device for malaria parasites in thick-blood films,” in Proc. IDAACS, Prague, Czech Republic, 2011, pp. 435-438. I. K. E. Purnama, F. Z. Rahmanti, and M. H. Purnomo.

BIOGRAPHY

Mrs.T.Grace Shalini working as Assistant Professor of Computer Science and Engineering Department, Velammal College of Engineering and Technology, Madurai-625009, Tamil Nadu.

M.Rajkumaris an UG student in the Computer Science and Engineering Department, Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

C.Arjun Rajis an UG student in the Computer Science and Engineering Department, Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

P.Suryais an UG student in the Computer Science and Engineering Department, Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

P.Arputharaj is an UG student in the Computer Science and Engineering Department, Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

.