

Malaria Detection Using Machine Learning with K nearest Neighbor Algorithm

Vallinayagam S¹ Thurun D² Kulothungan K³ Kolluri Mahesh K⁴ A.Velusamy⁵

^{1,2,3,4,5}Department of Computer Science and Engineering

^{1,2,3,4,5}Hindusthan Institute of Technology, Coimbatore, Tamilnadu India

Abstract— to detect the presence of Malaria parasites in Human Blood Cells and to determine how much it is affected. We have developed a machine learning method that can detect malaria parasites in thick blood smear images. Our method consists of two processing steps. To start with, 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. Malaria - a parasitic disease health problem which leads to millions of deaths especially in remote villages. This disease arises due to damaging of red blood cells. This paper presents a survey on detection / prediction of malaria disease using various machine learning techniques, Image Processing techniques and various clinical methods like rapid test, Nested PCR etc. In our observation, we found that machine learning techniques have wider applicability for critical diagnosis of malaria which in turn helps the clinicians for diagnosing the disease.

Keywords: Machine Learning

I. INTRODUCTION

An overview of detecting malaria disease based on data classification technologies, screening, pre-existing systems and their drawbacks, motivation behind the project, its objective and limitations, the functional and the non-functional requirements upon which the project is developed. This paper highlights the prediction of malaria diseases and the different factors responsible for it. It briefly describes the different data classification techniques in this work used for comparison.

II. PURPOSE

To design a tool to assist the physicians for malaria disease diagnosis. Motivated by the fast development of medical data classification, this work is designed to apply data classification technologies to detect early malaria disease prediction so that the relevant risk factors can be determined by individual by computer programs. The aim of data classification is to automatically or semi- automatically discover hidden knowledge, unexpected patterns and new rules from data.

III. SCOPE

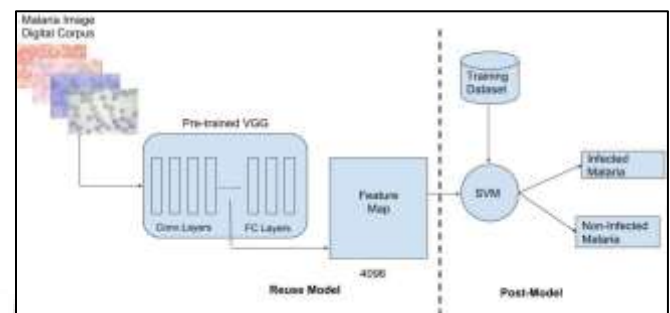
Malaria diseases depends on a lot of variables like hypertension, diabetes, exercise schedules, chest pain etc. This risk factors may not be scaled equally to every individual patient.

By the help of affected human grey scale image, the reports are given to the doctors for immediate actions to be taken.

Medical databases are usually filled by irrelevant and redundant features which increase the dimension of database or lead them to curse of dimensionality. It affects the accuracy, computational cost and speed of the learning process. □

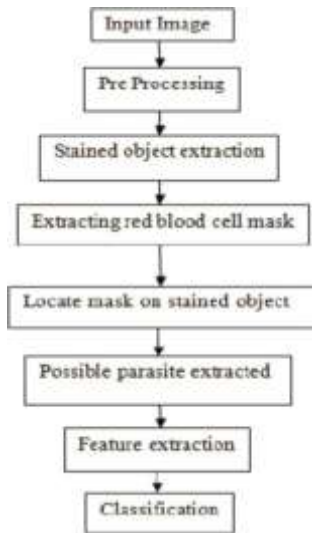
The accuracy and sensitivity have particular importance in the detection and prediction of diseases. Its positive feedback can predispose until the doctor by its analysis to speed up the process of diagnosis and prognosis

IV. ARCHITECTURE



A. 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 replaced 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 $44 \times 44 \times 3$ RGB patch image, with pixels having a distance greater than 22 to the center set to zero.



B. K-Nearest Neighbor Algorithm:

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until function evaluation.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k -NN classification) or the object property value (for k -NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

C. Malaria Disease Classifier (K-NN):

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 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 over fitting.

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, ResNe-50.

D. K-Nearest Neighbor Algorithm Steps:

- 1) Load the data
- 2) Initialize K to your chosen number of neighbors
- 3) for each example in the data
 - 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
- 4) Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- 5) Pick the first K entries from the sorted collection
- 6) Get the labels of the selected K entries
- 7) If regression, return the mean of the K labels
- 8) If classification, return the mode of the K labels

V. 2-D FEATURE EXTRACTION:

In this work, the front-end CNN is same as the first ten layers of VGG-16 with three pooling layers, considering the tradeoff between accuracy and the resource overhead. VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes. VGG16 was trained for weeks and was using NVIDIA Titan Black GPU's. There are 13 convolutional layers, 5 Max Pooling layers and 3 dense layers which sums up to 21 layers but only 16 weight layers. Conv 1 has number of filters as 64 while Conv 2 has 128 filters, Conv 3 has 256 filters while Conv 4 and Conv 5 has 512 filters. The image is passed through a stack of convolutional (conv.) layers, where the filters were used with a very small receptive field: 3×3 .

VI. SEGMENTATION BY DILATED KNN:

The back-end CNN is a series of dilated convolutional layers producing density map. The goal of semantic segmentation is classifying each pixel of the input image into a given set of classes. The main challenge of this is to combine pixel level accuracy with multi scale contextual information. The previous state of the art models are based on the adaptations of convolutional neural networks designed for image classification. The idea of Dilated Convolution was motivated as it enlarge receptive field while maintaining resolution and also it come from the wavelet decomposition. It is also called "atrous convolution" and "hole algorithm". This module is modified version of adapted VGG-16 network for semantic segmentation by removing the last two pooling and striding layers. Dilated convolutions utilize specific kernels with sparsely aligned weights. Both of the kernel size and the interval of sparse weights expand exponentially with dilation factor. By increasing dilation factor, receptive field is also expanded exponentially by large kernel. Dilated Convolution (Basic or Large) can always improve the results and does not overlap with any other post-processing steps. A dilated convolution is essentially a generalization of the

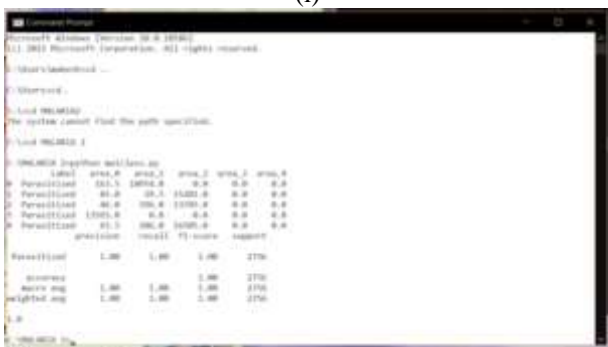
traditional 2D convolution that allows the operation to skip some inputs. This enables an increase in the size of the filter (i.e. the size of the receptive field) without losing resolution.

VII. OUTPUT USING K-NN ALGORITHM:

A good classification algorithm's accuracy must be high enough for better classification. Figure shows the accuracy of the classification results. K-NN enumerates the malaria disease diagnosis which consists of fully malaria related clinical data. Among the data record, K-NN provides better result with high accuracy. The attained accuracy shows the effectiveness of the MRF algorithm.



(i)



(ii)

A. Comparison of Accuracy between CNN and K-NN algorithm:

1) CNN Model

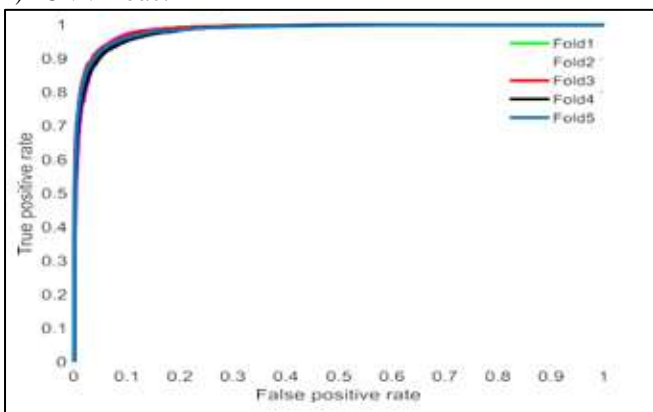


Fig. 1:

B. K-NN Model

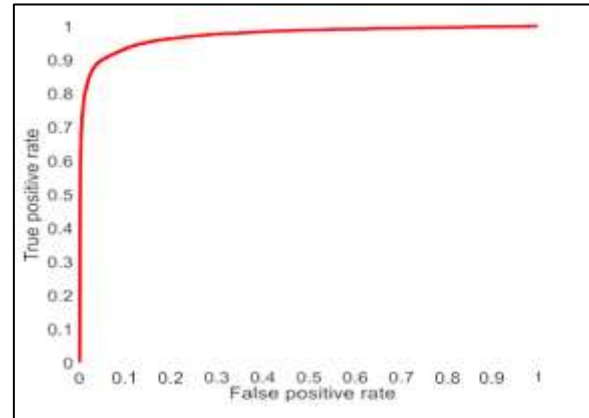


Fig. 2:

The ROC (Receiver Operating Characteristic) curve for both algorithm showed above. From Figure 1 and Figure 2, The proposed model (K-NN) showed the highest accuracy with highest true positive and true negative rate in the confusion matrix.

VIII. CONCLUSION

In this paper it is an overview of using Iterative Global Minimum Screening (IGMS) as a feature extraction with K-NN classifiers. It was observed that the performance rate of the classifiers after using IGMS to reduce the dimension of data improved. Experiments are conducted on 500 data records. Results show that using IGMS encouragingly enhances classification performance on most of the classifiers. From the graphs it is interpreted that the K-NN not only increases the support and confidence but also the run time taken is reduced considerably. Further implementation of the K-Nearest Neighbor (K-NN) will be generating the results which will be analysed.

REFERENCES

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.