PERFORMANCE COMPARISON OF MACHINE LEARNING ALGORITHMS FOR MALARIA DETECTION

Parvathi Pradeep,CB.EN.P2DSC21019

Abstract

Malaria is a life threatening disease that are transmitted to people by the plasmodium parasites that enter the body through the bites of the infected female Anopheles mosquitoes. It was, at one point, the major cause of death worldwide, but at present the disease is preventable and curable. The most common method of diagnosis is by observing the plasmodium in a peripheral blood smear sample. With the help of the images of these samples collected from the patients, the project aims to classify the image as that containing the infection or not using several classification algorithms and compare the classification metrics of each algorithm.

Index Terms

Malaria,Machine Learning,Haralick,Support Vector,Random Forest,Ada Boost, Decision Tree

I. Introduction

MALARIA can be prevented with early diagnosis and treatment which in turn reduces the transmission and deaths. WHO ‘s recommended methods of testing for it includes parasite-based diagnostic testing either with microscopy or a rapid diagnostic test. In the microscopic images of the blood samples, the aim is to find the presence of the plasmodium parasite. Several methods are used without the aid of an expert or trained personnel, to determine the presence by means of using computer vision and machine learning algorithms. With the help of image processing techniques, the image can be cleaned and processed for classification. In this project, the given images are preprocessed with the help of textural analysis and then provided as input to the machine learning algorithm to train the models and detect the type of cell, given a test image. The performance of each of these algorithms is then compared with the help of metrics such as the sensitivity-specificity trade off, accuracy and f1 scores.

A. Literature Review

The images used in the project are of poor quality and hence not much can be deduced from them, in their original quality. Not all of the information of the image is required to determine the infected cell. Certain distortions in the image, be it major or minor can be captured with the help of textural classification, that can reduce the feature dimensions of the image. Texture can provide information about the spatial arrangement of the intensity levels in the image. It can detect the patterns in variations of the image intensity. By using certain statistical methods on the gray level co-occurrence matrix (GLCM) of the image such as contrast, entropy and homogeneity, we can reduce the dimension of the image based on these measures. By using a set of 14 textural features as suggested by Haralick [1], we can capture maximum information about the image. The analysis of the poor-quality images taken from the standard microscope in [3] was done by first, labelling the images as infected or uninfected, making it as a binary classification problem. Then the necessary feature extraction was done using statistical representation of the shapes of the blood smears and not with the color as it is not informative with blood films. Then few machine learning algorithms such as KNN, Random Forest classifier, Ada Boost etc. were used to detect the parasitic cells. The conclusion was that Random Forest performed the best, with an accuracy of 0.965. The performance of the algorithms, based on the presence of parasite on the patch level, were compared with the help of classification metrics.

B. Objectives

In this case study, the aim is to classify the images as a parasitic or uninfected with the help of the classification algorithms and categorize each of their performances based on the metrics. The original features of the image will be replaced by the Haralick parameters obtained from each of the images, followed by the classification using the ml models. By comparing other classification metrics such as precision, recall and f1 score, we make sure that the performance of the model is just not based on the accuracy but also the capability of the model to reduce the number of misclassification on the whole, particularly the false negative cases.

C. Theoretical Background

The images taken in general consists of information stored in terms of pixel matrices where each pixel holds the intensity of the color of the image. When processing the images, it is tedious to consider all the features as input to the model. To reduce the load, a smaller set of features is extracted to represent the same image. The feature extraction process used in this project is texture analysis, that utilizes the intensities of the pixels of the images to differentiate cells with plasmodium from the ones without it.

Texture classification is identifying the textured region from a set of texture classes. The statistical approach to defining a texture has been explores in these cases where the texture is considered as the quantitative measure of the arrangement of the intensities in a region of an image. These measures can be considered as the feature vectors of the image. Statistical measures are useful to find the differences in texture at very small levels, which result in micro textures. This can be useful in the case study, where the parasite is depicted as spots on the image of the cell. With the help of the statistical measures, we can comprehend the spots which may not be visible to the naked eye, but will be detectable by the minute differences in intensities of the pixels in the image.To find any repeating nature or patterns in the texture of the image, the gray level co-occurrence matrix (GLCM) is used to obtain the positions of the pixels having the same intensity levels or gray levels. A co-occurrence matrix will have the size n*n where n will be the number of gray levels, n, considered to represent the gray scale image. An entry in the matrix G, G[i, j] is defined by finding the pair of pixels separated by a given displacement vector d having the gray cells i ,j. We normalize this matrix by dividing each entry with the total number of pixel pairs so this can be thought of as probabilities too. Some of the statistics that can be derived from the normalized matrix are maximum probability, moments, contrast, homogeneity of the image and so on.
Haralick features are a set of 14 textural features that can be extracted from the normalized gray scale co-occurrence matrix which gives the information about the image's textural characteristics such as homogeneity, linearity and contrast. In this case study, the 14th feature vector is not calculated due to high computation time. Thus, the images are converted from their original size to a 13-dimension feature vector for the analysis. In addition to these 13 features, the Hu moments at 7 threshold levels have been included in order to capture the representation of the shape of the plasmodium from all images, which will be similar irrespective of the translation, scaling or rotation of the images.

The classification algorithms used includes Ada Boost Classifier which runs on boosting the given base model, Random Forest Classifier that performs bagging, primarily on the Decision tree Classifier, K Nearest Neighbors, probability base Naive Bayes Classification, Support Vector Classifiers and Logistic Regression.

D. Methodology

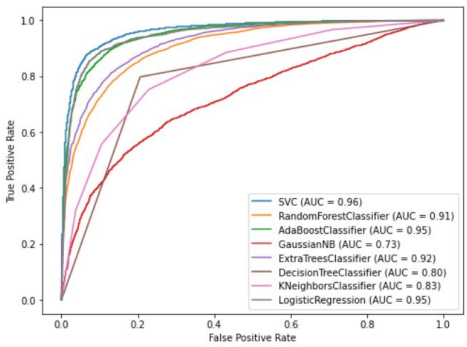
The image data set used for this project if from the collection of malaria screen-er research from Lister Hill National Center for Biomedical Communications [2]. These images are of very low resolution and are difficult to label as parasitic or uninfected at a first glance as they are developed by using a conventional light microscope aided with a smartphone camera. There are a total of 27558 instances taken with equal number of parasitized and uninfected cell images. After retrieving the data from their respective files, the data is then individually labelled as parasitic or uninfected depending on the folder they were stored in. Then the data set was split for training and testing, where 75% of instances were for training and the remaining were used to test the models.

Each of the images is then subjected to a function that takes in the image and then processes it's haralick features for feature reduction. This is done with the help of the mahotas package. The original features of the gray scale image are extracted with the help of the cv2 package and then these are fed as the input to the haralick features extractor. The features as output are the 13 statistical methods of texture analysis identified for each of the images. The moments were also calculated with the help of the Hu moments function. The function appends the labels along with the features to complete the data set for the model to classify on.

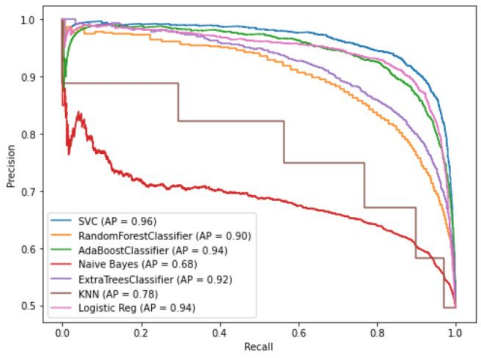
In this scenario, we have labelled 0 for the uninfected cell images and 1 for the parasitic cell images and hence, is a binary classification problem. The machine learning algorithms used include the Ada Boost Classifier, Random Forest Classifier, K Nearest Neighbors, Naive Bayes Classification, Support Vector Classifiers and Logistic Regression. The parameters and hyper parameters of these models have been adjusted accordingly to maximize the probability of correct predictions. Support Vector Classifier algorithm is seen as the best performing algorithm based on the accuracy with a score of 0.91. The prediction is based on the parasite detection on the images of the cell and does not necessarily extend to the condition of the patient.

E. Experimentation Results and Discussions

The images taken from https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html#malaria-datasets are separates into 75% training set i.e., out of 27558 samples, 20668 samples containing both parasitic and uninfected cell images were taken to train



(a) ROC Curve



(b) Precision Recall Curve

the models and the remaining 6890 cell images were used for testing.

The figures 1 and 2 show the ROC (Receiver Characteristic Operating) and Precision recall curves taken for all the considered algorithms. From the ROC curve, the Support Vector Classifier has the highest area under the curve in the ROC curve graph with a value of 0.96.In the Precision recall graph,the average precision score (AP) is highest for Support vectors with a value of 0.96. This conveys that at 96% precision, or at high precision in general, lesser number of false negatives (between 10-20%) will be observed.

The table below summarizes all the metrics taken for the considered algorithms.The Support Vector Classifier gives the highest accuracy of 0.91, followed by Logistic Regression and Ada Boost algorithms.

Classification metrics				
Machine Learning Algorithms	Accuracy	Precision	Recall	F1 score
Support Vector Classifier	0.9062	0.9062	0.9062	0.9062
Random Forest Classifier	0.8284	0.8287	0.8285	0.8284
AdaBoost classifier	0.8842	0.8842	0.8842	0.8842
Naive Bayes	0.6517	0.6565	0.6511	0.6584
Extra Trees Classifier	0.8433	0.8436	0.8434	0.8432
KNN	0.7573	0.7574	0.7574	0.7573
Logistic Regression	0.8922	0.8924	0.8921	0.8921
Decision Tree Classifier	0.7929	0.7929	0.7929	0.7929

II. Conclusion

This project discusses the detection of malaria parasite with the help of textural classification to reduce the feature dimension of the collected images and the classic machine learning algorithms to predict the images as an uninfected or parasitic cell image. For the feature extraction, we have used the Haralick features and the moments from the image. The algorithms used for the classification include Support Vector classifiers, Random Forest, Ada Boost, Decision Trees, etc.The textural classification can be effective in cases where the presence of a foreign body in a cell are able to be identified by the difference in the intensities of pixels in the image. The accuracies can be further improved by use of neural networks and deep learning techniques.

References

[1] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[2] Sivaramakrishnan Rajaraman, Sameer K. Antani, Kamolrat Silamut Mahdieh Poostchi, Md. A. Hossain, Stefan Jaeger Richard J. Maude, and George R Thoma. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. apr 2018.

[3] G Saiprasath, Naren Babu, J ArunPriyan, R Vinayakumar, V Sowmya, and K Soman. Performance comparison of machine learning algorithms for malaria detection using microscopic images. *IJRAR19RP014 Int. J. Res. Anal. Rev.(IJRAR)*, 6(1), 2019.