



e-ISSN: 2319-8753 | p-ISSN: 2320-6710

# IJIRSET

International Journal of Innovative Research in  
**SCIENCE | ENGINEERING | TECHNOLOGY**



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

Impact Factor: 7.569



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

# Malaria Detection Using Supervised Learning

Prof.S.T.Sawle<sup>1</sup>, Samruddhi Sudhir Kavimandan<sup>2</sup>, Prachi Pradip Narkhede<sup>3</sup>

Associate Professor, Department of Information Technology, Anuradha Engineering College, Chikhli.

Maharashtra, India<sup>1</sup>

U.G. Student, Department of Information Technology, Anuradha Engineering College, Chikhli, Maharashtra, India<sup>2,3</sup>

**ABSTRACT:** Malaria is a deadly, infectious and life threatening mosquito-borne blood disease caused by Plasmodium parasites. The traditional and most standard way to diagnose malaria is to examine the blood smear under a microscope. This method is inefficient and time consuming and depends on the diagnostic experience and knowledge of the person taking the test. Automated image recognition technology based on image processing has previously been applied to malaria blood smears for diagnosis. However, practical performance has not been limited so far. It inspires all of us to make the diagnosis and diagnosis of malaria faster, easier and more efficient. Our main goal is to create a model that can detect cells from multiple cell images of a thin blood smear on a standard microscope slide and classify them as infected by early or effective testing using an image process. And classify infected cell images using machine learning. Keywords: Malaria, Falciparum, Watershed, Morphological Segmentation, Age Detection, Segmentation.

## I.INTRODUCTION

Malaria is a deadly and contagious disease caused by the Plasmodium parasite, which is transmitted by the bite of the female Anopheles mosquito. According to the 2012 World Malaria Report published by the WHO [1], an estimated 100,000 malaria-related deaths were recorded last year. The disease can be cured but early diagnosis is the main thing. Existing methods used to detect malaria include microscopic examination of infected cells in the laboratory. This method is both expensive and tedious. In 2012, the WHO reported about 100 percent of all malaria cases in Africa. The region has one of the highest per capita incomes in the world. This model offers a fast, low cost and reliable alternative to malaria micro-testing.

**i.Problem Statement :** We propose an image processing model to detect malaria infected cells. We use image processing techniques to detect parasitic infected red blood cells in a thin smear on standard microscope slides. The most commonly used method nowadays is to analyze thin blood smears under a microscope and look at the perspective of infected cells. A physician himself counts the number of parasitic red blood cells - sometimes up to 1000 cells (according to WHO protocol) [2].

If increasingly accurate and effective symptoms are accessible to the technique, malaria can be deforested, controlled and eradicated everywhere. We have used image processing to identify near malaria contaminated cells. And to classify the condition of malaria, whether it is falciparum or not, it is the most deadly stage in malaria or non-falciparum, for which we use machine learning technology.

**ii. Scope of the Project:** Malaria screening from thin film blood smear images demands the separation of single blood cells from microscopic blood slide images that can be taken from a pathologist and the dataset does not have cell images that do not have division. Therefore, in the proposed method the partitioning is done using different image processing techniques. The edge detection technique and segmentation technique used in this system detect cell discrepancies by removing cells by sound. It differentiates between cells and identifies infection in cells using morphological segmentation. Also, all the images are raw and their intensity is different and it is very difficult to detect cells and infection as there is no uniformity in all the images. To overcome this problem, the proposed method uses histogram matching where all images are standardized and the same intensity also increases the level of accuracy.

## Types of Machine learning algorithm

### 1. Supervised Learning

**How it works:** This algorithm contains a target / result variable (or dependent variable) that must be inferred from a set of predictive variables. Using a set of these variables, we generate a function that maps the desired output. The

training process continues until the desired accuracy level on the model training data is achieved. Examples of supervised education: regression, decision tree, random forest, KNN, logistics regression etc..

## 2. Unsupervised Learning

**How it works:** In this algorithm, we have no target or result variable to predict / predict. It is used for population clustering between different groups, which is widely used to divide customers into different groups for specific interventions. Examples of ineffective learning: ri priori algorithm, k-means.

## II.LITERATURE SURVEY

The Plasmodium parasite is a parasitic malaria that invades red blood cells (RBCs) and is transmitted by mosquitoes. Neurol networks are used to analyze the potential of RBC and parasitic ethin blood smears. In, the loaded KNN (K-close neighbor) algorithm is trained by the options taught in the Abuse Theorem Picture Component Classifier, which aims to spot pixels. To identify multicategory parasites according to the lifecycle stage and its species



Fig1 Worldwide malaria death rates

Basic thresholding is accomplished using a bar graph-based procedure to detect the presence of plasmodium in blood smears. It is very important to prepare the smear, as these changes can lead to changes in imaging conditions. The misbehavior of the overlapping RBC was the morphological operator. Deformed square measurement reported by analyzing cells containing cells where the image of authenticity is binaryized and using a vague live technique. Further, these tagged cells have some properties such as color, size, and choice as a rank neural network mistreatment using square-sized platelets, leukocytes, and corpuscles classified designs. Over the past few years there has been a good deal of agreement to develop new methods for detecting protozoal infections, including fast substances, fluorescent research detection methodology, and PCR (polymerase).

Chain reaction) The process of finding specific sequences of macromolecules. Nonetheless, the method of identifying light weight research is the most common and commonly used technique.

The research will differentiate between species categories, quantify parasitism, and monitor the different acute stages of the parasite. But this method requires trained technicians and it is a time consuming process and the ultimate precision of identification also depends on the skills and abilities of the scientists and the amount of time they spend learning each slide. Malaria is a life-threatening disease and scientists around the world are interested in analyzing it. Previously, a large number of protozoal infections were diagnosed in a laboratory setting that required good human experience. Automated systems such as hope for machine learning techniques have previously been studied to overcome this negativity. During the study of this domain, technology gave a lot of thought to hand-crafted options when reporting. For example, SVM and Principal Part Anal Analysis (PCA) were applied for morphological factors and classification purposes for characterization. However, the accuracy achieved by these types of models is less than that of the in-depth learning-based techniques studied recently.

What is special is that during this work, we have a tendency to propose in-depth models that bring home the Bacon classification performance, such as the highly accurate deep learning already reported. In addition, our models measure economically in terms of required processing resources and fail to perform efficiently on a good mobile device, as well as with the square measures offered at a low cost. [3]

### III.MACHINE LEARNING TECHNIQUES

#### A.Supervised Learning

Supervised education is the most common sub-branch of machine learning today. In short, new machine learning practitioners will begin their journey with supervised learning algorithms. So, the first supervision in this three post series will be about education.

Supervised machine learning algorithms are designed for example learning. The name "supervised" education comes from the idea of training this type of algorithm as teachers supervise the whole process.

When training supervised learning algorithms, the training data will include inputs connected with the appropriate output. During the training, the algorithm will search for patterns in the data related to the desired output. After training a supervisory learning algorithm will take new unseen input and determine which labels will be classified based on previous training data. The purpose of the supervised learning model is to predict the appropriate label for the newly introduced input data. In its most basic form, supervisory learning algorithms can simply be written as:

$$Y=f(x)$$

Where y is the predicted output that is determined by the mapping function that assigns the square to the input value x. The task used to connect the input features to a prediction is created during training by a machine learning model. Supervised learning can be split into two subcategories[10]:

#### Classification and regression.

##### a.Classification

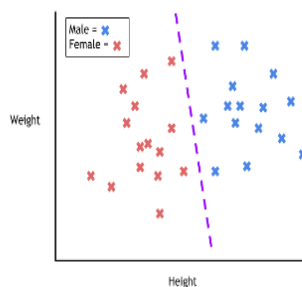


Fig2

During training, a classification algorithm will be given data points with an assigned category. The function of the classification algorithm is to then take the input value and assign it a class or category based on the training statistics provided to it. The most common example of classification is determining whether or not email is spam. With two classes to choose from (spam or no spam), this problem is called binary classification problem. The algorithm will be provided with training data containing non-spam and non-spam emails. The model will find features in the data that are either consistent with the range and create the mapping function mentioned earlier:  $Y = F(X)$ . Then, when an unseen email is provided, the model will work to determine if the function is spam.

Classification problems can be solved with many algorithms. Which algorithm you choose to use depends on the data and the situation. Here are some popular classification algorithms:

- Support Vector Machines
- Decision Trees
- K-Nearest Neighbor
- Random Forest

##### b Regression

Regression is a statistical process of prediction where the model seeks to find a significant relationship between dependent and independent variables. The goal of the regression algorithm is to predict continuous numbers such as sales, revenue, and test scores. The basic linear regression equation can be written as:

$$Y=w[0]*x[0]+w[1]*x[1]+...+w[i]*x[i]$$



Where  $x[i]$  is the feature(s) for the data[7] and where  $w[i]$  and  $b$  are parameters which are developed during training. For a simple linear regression model with only one feature in the data, the formula looks like this:

$$\hat{y} = wx + b$$

Where  $W$  is the slope,  $x$  is the only feature and  $B$  is the y-intercept. Familiar? For simple regression problems like this, the model estimates are indicated by the best fit line. This aircraft will be used for models that use two features. Finally, hyperplanes will be used for models that use more than two features. Imagine being able to determine a student's test grade based on how many hours we studied during the exam week. Let the plotted data with the best fit line look like

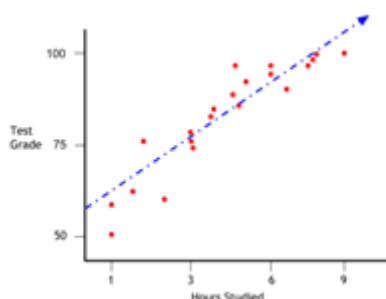


Fig3

There is a clear positive relationship between the hours studied (independent variables) and students' final test scores (dependent variables). Given the new input, the best fit line can be drawn through the data points to show the forecast of the models. Say we wanted to know how well a student would do with five hours of study. We can use the best fit line to estimate test scores based on other students' performance.

There are various types of regression algorithms. The three most common are listed below:

- Linear Regression
- Logistic Regression
- Polynomial Regression

## B.Unsupervised Learning

You only have insecure learning without input data ( $x$ ) and corresponding output variables. The purpose of unsupervised learning is to model the underlying structure or distribution in the data to learn more about the data. This is called unsupervised learning because without the above supervised education there are no correct answers and no teacher. Algorithms have been released on their own devices to find and present interesting structures in the data. The problem of insecure learning can be further divided into the problem of clustering and association.[1]

### i.Clustering:

A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior[1]. Clustering is a data mining technique that groups unlabelled data based on their similarities or differences. Clustering algorithms are used to group information on raw, unclassified data objects by structures or patterns[6]. Clustering algorithms can be classified into several types, especially unique, overlapping, hierarchical, and potential.

### ii.Association:

The problem with learning association rules is where you want to find rules that describe a large portion of your data, just as people who buy  $X$  also buy  $y$ [4]. An association rule is a rule-based method for finding the relationship between variables in a given dataset. . These methods are frequently used for market basket analysis so that companies can better understand the relationship between different products. Understanding customer usage habits enables businesses

to develop better cross-selling strategies and recommendation engines. Examples of this can be found in Amazon's "Customers Who Purchased These Items" or in Spotify's "Discover Weekly" playlist. While some different algorithms are used to generate association rules such as apriori, eklat, and fp-growth, the ri priori algorithm is the most widely used.

#### IV.RANDOM FOREST IN MACHINE LEARNING

Random forest algorithm is most famous and easy to use machine learning algorithm based on ensemble learning .In this article you will learn how this algorithm works,how it's efffecient comparing to the other algorithms.

##### A.What is Random Forest in Machine Learning?

Random Forest is a supervised machine learning algorithm that can be used to solve classification and repression problems[6]. However, most of them prefer classification .This is called a random forest because it connects multiple decision trees to create a "forest" and feeds them random features from the provided dataset. Instead of relying on individual decision making, the random jungle estimated all the trees and selected the best result through the voting process.

Now the question arises as to why we have preferred random forests over decision making. So, individual seedlings are more useful but random forests can eliminate this problem by averaging the approximate result on each plant.

The basic idea of random forestation should be explained here. Let's dive deeper into this and understand how this algorithm works.

##### How it works?

We can understand the working of a random forest with the help of the given steps:-

- Select random patterns from the dataset provided by Data.
- Selected Create a decision tree for each selected pattern. You will then get approximate values for each tree created.
- There will then be a vote for each prediction test
- In the end, the algorithm will choose the result(predicted) with majority votes

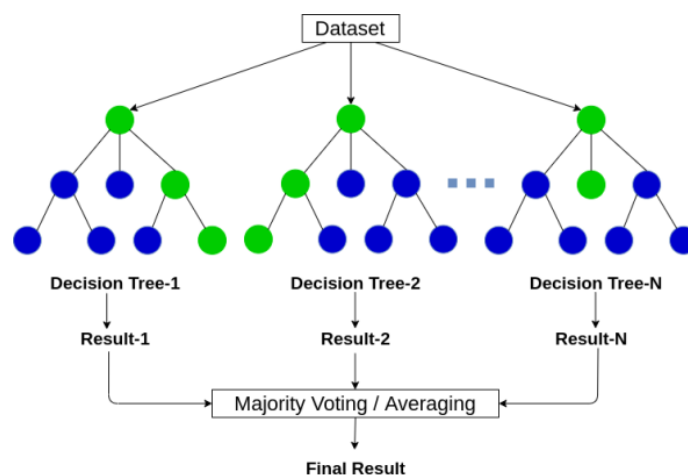


Fig 4 Working in Random forest

## V. ADVANTAGES AND DISADVANTAGES

### A. Supervised Learning

#### Advantages

1 The example of linear regression is simple and straightforward to understand. This can be normal so that they do not feel uncomfortable. Furthermore, linear models with new data can be easily updated using statistical gradient

2 The use of well-known and labeled input data makes supervised learning much more accurate and reliable than obsolete learning. By accessing labels, it can be used to improve its performance on specific tasks.

3 Efficient in finding solutions to several linear and non-linear problems such as classification, robotics, prediction and factory control. Neurons able to solve complex problems by hiding leather (Satya and Abraham, 2013).

#### B. Disadvantages:

1 Takes a long time for the algorithm to compute by training because supervised learning can grow in complexity. Therefore, since most of the data in the world is unlabeled, results are not actually obtained, performance is very limited.

2 Performs poorly when there are non-linear relationships. A method of supervisory learning such as linear regression that is not flexible to capture more complex structures. It takes a long time to calculate and it is also difficult to add the right polynomial or interaction terms..

3 Its not cost efficient if the data keeps growing that adds to the uncertainty of data labelling to predefine outputs. Example, It is costly to manually label an image dataset, and the most high quality image dataset has only one thousand labels, according to (Ankur A., 2018).

### B. Unsupervised Algorithm

#### Advantages:

1 Lets algorithm to refer back for patterns that has not been accounted previously[11], therefore resulting the freedom of learning direction for the algorithm in unsupervised learning (Kohonen and Simula, 1996).

2 Inadequately labeled dataset or unknown pattern to identify or violate the problem while constantly evolving. Learning the hidden pattern of data trained on it. Previously managed problems become more solvable and more agile in finding hidden structures in past data and future predictions (Ankur A., 2018).

3 Simplified human task of labelling by grouping similar object and differentiating the rest. This group of datasets will then be labeled one after the other instead of being labeled (Ankur A., 2018).

#### Disadvantages:

1 (Stuart & Peter, 1 1996 1996) According to him, a learned person who does not have complete information cannot learn what to do in certain situations as the information is not provided. The goal of insecure learning, which is not accessible in any output, is simply to find a pattern in the available data feed..

2 Quite slow and consumes large resource memory, therefore harder to scale to larger datasets. Furthermore, it assumes that only the original clusters in the dataset are globe-shaped.

3 The outcomes are not that accurate due to it is mostly about prediction. . In addition, we do not know the number of classes, so the result is not certain. Unsupervised education is less adept at solving a narrowly defined problem (Sylvia, 2018).

## VI.RESULT

Here Random Forest algorithm used in our topic because it has more accuracy than others.

See the table below

Algorithm	Accuracy	Precision	Recall	F-Score
Decision Tree	0.965	0.526	0.529	0.527
KNN	0.940	0.465	0.278	0.348
Linear Regression	0.943	0.375	0.000	0.000
Naive Bayes	0.858	0.271	0.873	0.414
Random Forest	0.965	0.775	0.553	0.645
Extra Tress	0.956	0.837	0.298	0.440

Performance of different Machine learning algorithm

Python has various in-built libraries used specifically for implementing machine learning algorithms. Fortunately, we don't have to write code to implement complex parts, we can accomplish things just by importing these libraries. We use the Random Forest Classifier algorithm here because it has more accuracy than other algorithms.

## VII.CONCLUSION

We have proposed a Malaria parasite detection method using a shallow machine learning algorithms. This method of detecting the malaria parasite can be very useful to health workers in countries, where there is less number of trained laboratory experts and lack of resources. In the present work, we divided the image into patches and analysed based on the presence or absence of malarial parasite. To accomplish this, we have used various classical machine learning algorithms such as AdaBoost, Decision Tree, KNN, Random Forest, etc. The accuracy of our model assists the laboratory technicians in decision and classification framework may be sufficiently general for other diagnostic tests like hemoparasites, worm infestations, or tuberculosis

## REFERENCES

- [1] World Health Organization, Malaria, <https://www.who.int/newsroom/factsheets/detail/malaria-report-2019>
- [2] World Health Organization, Malaria, [https://www.who.int/newsroom/factsheets/detail/malaria\(2018\)](https://www.who.int/newsroom/factsheets/detail/malaria(2018))





**Impact Factor:  
7.569**



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



[www.ijirset.com](http://www.ijirset.com)

Scan to save the contact details