# Malaria Disease Prediction Using AI

Ahmed Emad[1], Mohamed Said[2],
Kirollos Emad[3], Marwan Fathy[4]
*Faculty of Computer Science*
*Misr International University, Cairo, Egypt*
Ahmed2011640[1], Mohamed2010565[2],
Kirollos2010269[3],Marwan2014606[4]{@miuegypt.edu.eg}

***Abstract*—Malaria is a global health issue that affects millions of people worldwide. The traditional method of diagnosing malaria involves manually examining blood samples under a microscope to detect the presence of the malaria parasite. This process can be time-consuming and prone to errors. In this paper, we will introduce malaria and how it can traverse from one human to another then we will discuss how it can be detected using technology based methods. We will also review previous research on machine learning and deep learning models for detecting malaria. Our proposed methodology involves using various types of classifiers to detect malaria. We will share the dataset we used to train these models and present the results of our experiments in a table, comparing the accuracy of each model. Overall, our goal is to provide a comprehensive overview of the use of machine learning and deep learning techniques for detecting malaria and to present our own approach and results in this area.**

## I. Introduction

Malaria is a serious and potentially life-threatening disease caused by parasites transmitted through the bites of infected mosquitoes. It is a major public health problem, particularly in tropical and subtropical regions, where it is a leading cause of illness and death [1]. According to the World Health Organization (WHO), there were an estimated 229 million cases of malaria worldwide in 2019, with 405,000 deaths, mostly among children under the age of 5 in the African region [2] [3]

The burden of malaria varies widely among countries and regions. In 2019, the highest number of cases was reported in the WHO African region, followed by the South-East Asia and the Western Pacific regions . However, the highest incidence of malaria (the number of new cases per 1,000 population) was reported in the WHO African region, followed by the Eastern Mediterranean and South-East Asia regions [2]

Efforts to control and eliminate malaria have made significant progress in recent years. The use of effective prevention and control measures, such as insecticide-treated bed nets, indoor residual spraying, and antimalarial drugs, has contributed to a global decline in malaria cases and deaths [2]. However, progress has been uneven, and there are still many challenges to overcome in the fight against malaria

Continued efforts are needed to improve the availability and quality of diagnostic and treatment services, strengthen surveillance and response systems, and develop and scale up new prevention and control measures [4] Collaboration between governments, international organizations, and the private sector is also crucial in addressing the complex and multifaceted nature of the malaria problem (Payne et al., 1988; Warhurst et al., 1996).

Malaria is caused by parasites of the Plasmodium species, which are transmitted to humans through the bites of infected mosquitoes. When a mosquito carrying the parasite bites a person, the parasite is introduced into the person's bloodstream. The parasites then travel to the liver, where they multiply and eventually enter red blood cells. Infected red blood cells don't seem enlarged in P. knowlesi. P. knowlesi parasites have an identical morphology to P. malariae, however their living substance is additional irregular. Trophozoites could have protozoal infection pigment unfold inside, a band kind almost like P. malariae, and multiple parasites will infect one red blood cell.[4]

Leishman' stain (1901) is a technique that was used to detect malaria by viewing microspy blood sample. This technique encompasses a high sensitivity, is inexpensive, and within reason straightforward to use. The Wright-Giemsa stain, for example, is a mixture of Wright and Giemsa stains, with the previous facilitating the excellence of blood corpuscle types. [4]

Giemsas stain (1902) was used for the primary time to diagnose protozoal infection quite a century ago. This method gotten loads of attention since then. it's presently ordinarily utilized in microscopical malaria investigations as a result of to its cheap cost, glorious sensitivity, and specificity. Giemsa staining necessitates several chemicals, masterly personnel, requires lots of effort ,and takes alot of time (it generally needs a minimum of forty-five minutes to stain a slide).[5] alternative stains are utilized as well, admire Field stain, that greatly decreases staining time whereas requiring sample drying before and through staining. Field' stain, on the other hand, has drawbacks, notably at low-resource health establishments wherever it's going to be

applied. Bacteria, fungus, stain precipitation, dirt, and cell rubbish are all frequent artifacts misinterpreted for protozoal infection parasites as a results of poor blood preparations. False-positive readings are common as a result of them.[6]

Sodeman et al. studied the impact of dye staining in characteristic protozoal infection parasites at low levels of infection within the 1970s. Dye staining has been incontestible to be additional sensitive and long than Romanowsky and Giemsa staining procedures, however it wants plenty of experience and training, and it's artifacts like photobleaching and phototoxicity. Furthermore, light microscopes are costlier than traditional light microscopes, that may be a thought in resource-scarce tropical areas wherever malaria is prevalent.[4]

Artificial intelligence (AI) has the potential to improve the accuracy and efficiency of malaria diagnosis. For example, AI algorithms can be used to analyze microscopy images and accurately identify malaria parasites [4]. AI algorithms can also be trained to recognize malaria parasites in rapid diagnostic test (RDT) results (Owusu-Agyei et al., 2019). We're going to see the results of those cell images by our system, also analyze all the possible test given by doctors discussing the malaria disease. All this information will be collected on a data set that is going to be used in numerous probabilistic models made to predict malaria.

In recent years, infection illness has become an extremely cogent topic due to its significance on the worldwide health. Furthermore, analysis on machine learning has targeted heavily on the protozoal infection prediction, driven largely by the discharge of giant footage datasets like the protozoal infection detection dataset through the event of the latest feature extraction and learning ways. Current datasets contain tough images. The performance of varied ways on such datasets has additional and more improved over the past several years.

Because the number of features extracted in machine learning training instances is fixed, the number of samples and multivariate variables required grows exponentially, and the classifier's performance suffers as a result. As a result, the algorithm slows or stops learning. This curse is overcome by using lower-dimensional space accurate sampling/ mapping to compensate for the loss of information in rejected variables.we're going to face these difficulties in the machine learning and try to predict this disease. Malaria disease prediction can save millions of lives all over the globe.[7]

The major contributions of this paper after feature extraction methods are as follows:

- SVM.
- KNN.
- Descision Tree.
- Random Forest
- Naive Bayes

The remainder of this paper is organized as follows. Section 2 discusses related work. In Section 3, we have a tendency to describe our methodology for this study. In Section 4, we concisely discuss varied feature extraction approach and algorithms, reminiscent of SVM,KNN,Descision Tree. Section five provides a comprehensive analysis of our approach on the algorithms. Finally, our conclusions and future work.

## II. Related work

The study "Deep Learning-Based Malaria Detection in Blood Smear Images" published on the arXiv preprint server in 2020 describes a system that uses deep learning to detect the presence of malaria parasites in blood smear images.

The study begins by introducing the problem of malaria, including its prevalence and the importance of accurate and timely diagnosis. The authors then describe the current methods for diagnosing malaria, including microscopy and rapid diagnostic tests, and discuss the limitations of these methods.

To address these limitations, the authors propose a deep learning-based approach for automated detection of malaria parasites in blood smear images. The deep learning-based approach consists of a convolutional neural network (CNN) that was trained on a dataset of images to recognize the presence of malaria parasites.

The authors preprocessed the images by resizing them to a fixed size and applying histogram equalization to improve contrast. They then split the dataset into training, validation, and test sets, with the training set comprising approximately 70 percent of the images, the validation set comprising approximately 15 percent of the images, and the test set comprising approximately 15 percent of the images.

The authors experimented with several different network architectures for the CNN, including a simple feedforward network and more complex architectures such as Inception and ResNet. They also finetuned the hyperparameters of the CNN, such as the learning rate, the batch size, and the number of epochs, to optimize the performance of the network. [8]

In the study"Classification of Malaria Using Object Detection Models" Roy et al. (2018) detected the malaria parasite in the Giesma blood sample using image processing, where they developed a model that used the color pixel-based discrimination method and a segmentation operation to identify malarial parasites in microscopic blood smear images [9].

Their methodology involved using two different segmentations: watershed segmentation and HSV (hue,

saturation, and value space) segmentation. Then they followed the morphological operations to highlight the presence of the parasite in the microscopic images of RBCs. Their methods resulted in a 90 percent accuracy in the detection of the parasite causing the disease.

Scherr et al. (2016) proposed a method to analyze malaria using mobile phone imaging and cloud-based analysis for standardized malaria detection, where a mobile phone was used to take images for conducting a rapid diagnostic test.

It also enabled the objective recording of the rapid diagnostic test, and this enabled web access for immediate result reporting.

These images were uploaded to a database that was globally accessible and then analyzed.

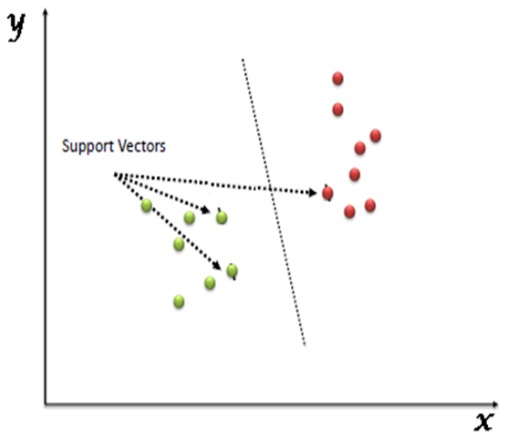
It achieved an 80.2percent true negative rate (specificity). This was a novel use of digital pathology to ensure top-notch healthcare for patients [10]. [11]

## III. Proposed Methodology

This portion describes the main modules of the suggested framework for predicting malaria sickness. Which includes feature extraction, image enhancement, and classification using different types of classifiers with different algorithms which are (KNN, DecisionTree, RandomForest, SupportVectorMachine (SVM), Naïve Bayes). Firstly, we select all the training photos from the dataset and place them in one array. After that we divide the array into two arrays (data and labels) where we extract all the data from the photos. Furthermore, after the data had been extracted, we convert the data into usable data inform of binary data (0,1) and flatten them so that we can pass each of them to the classifiers to be trained. Finally, we check the results of our classifiers. which are (Training set, validation set and test set) will provide CA (classification accuracy), F1(harmonic mean), Precision, and Recall are the outcomes of this method.

### A. SVM

Support vector machines are a group of supervised learning techniques for classifying data, doing regression analysis, and identifying outliers. When using the SVM algorithm, each data point is represented as a point in n-dimensional space (where n is the number of features you have), with each feature's value being the value of a certain coordinate. Next, we perform classification by identifying the hyper-plane that effectively distinguishes the two classes. There are two types of SVM simple and Kernel. We used Kernel SVM in our project as it Has more flexibility for non-linear data because you can add more features to fit a hyperplane instead of a two-dimensional space.




### B. KNN

The result of k-NN classification is a class membership. The class that an object is assigned to base on the majority vote of its k closest neighbors is determined by the item's neighbors' output of k-NN regression is the object's property value. The average of the values of the k closest neighbors determines this value. Firstly, we noticed that the KNN take short time in the training area while taking a longer time while being in the testing phase. in addition to that by decreasing the number of k closest neighbors the accuracy of the testing would increase

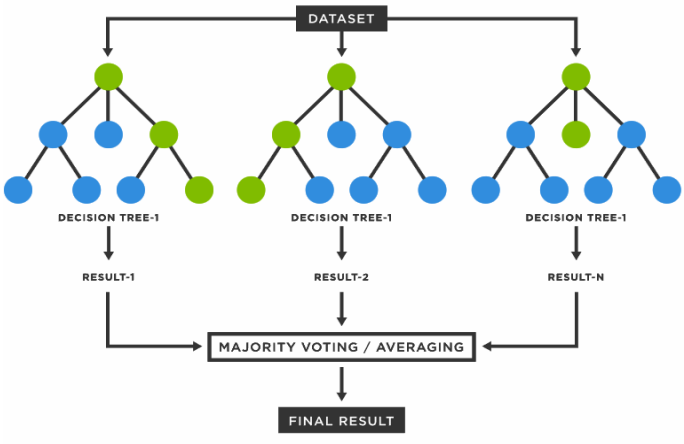
$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

### C. Decision Tree

A decision tree is a flowchart-like structure in which each leaf node represents a class label, each internal node represents a "test" on an attribute, and each branch indicates the result of the test (decision taken after computing all attributes). Classification rules are represented by the routes from root to leaf. The challenging part that faced us during our testing was trying to increase the accuracy of the classifier without overfitting the data by increasing the depth of the tree which will lead to increasing the variance

### D. Random Forest

Random forest classifier operates by building multiple decision trees during training phase. The class that most of the trees choose is the output of the random forest for classification problems in the testing phase, The mean or average prediction of each individual tree is returned for regression tasks. The tendency of decision trees to overfit their training set is corrected by random decision forests. Random forests generally outperform decision trees.




### *E.* Naive Bayes

Naïve Bayes algorithm is a supervised learning algorithm, it is a probabilistic classifier, which means it predicts based on the probability of an object. which is based on Bayes theorem and used for solving classification problems.

$\mathrm{P(A \mid B)} = \frac{P(B|A)P(A)}{P(B)}$

In our testing studies we used two types of Naïve Bayes algorithm 1-Gaussian Model: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

$P\left(x_i \mid y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$

2- Multinomial Model: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.The classifier uses the frequency of features for the predictors.

## IV. Experimental results

We test our technique on many activities from provided public dataset in this section.

### *A.* malaria disease datasets

We like to base our approach to the check in this section on a variety of actions from two open Data sets. The two categories of the project are "Parasitized" and "Uninfected." Our data is split into two sections: Training and Testing. There are 15,832 photos within the testing dataset, that are separated into two categories: Parasitized and Uninfected. There are 7,952 parasite photos and 7,880 uninfected pictures in the testing data set . The training dataset, on the other hand, has 27,560 photos divided into 13,780 Parasitized and 13,780 uninfected images. The classifier was trained on the training set, while the validation set was used to fine-tune the parameters for every technique.

### *B.* Performance metrics

precision (PREC), recall (REC), and f1 were all used to evaluate the suggested approach. The parameters for evaluating various strategies are defined based on the metrics. In addition, we show the best result's confusion matrix for each dataset. In this project we used several methodologies like KNN, SVM , Decision Tree,Random Forest and Naive Bayes with our main goal is to get the best performance

TABLE I
PERFORMANCE MEASURES USING THE CLASSIFIERS FOR PARASITIZED TEST

| Classifiers | AUC | F1 | Precision | Recall |
|---|---|---|---|---|
| SVM | 0.86 | 0.86 | 0.86 | 0.87 |
| KNN | 0.77 | 0.74 | 0.88 | 0.64 |
| Decision Tree | 0.74 | 0.72 | 0.78 | 0.66 |
| Random Forest | 0.99 | 0.99 | 0.98 | 1.00 |
| Naive Bayes | 0.66 | 0.64 | 0.68 | 0.60 |

TABLE II
PERFORMANCE MEASURES USING THE CLASSIFIERS FOR UNINFECTED TEST

| Classifiers | AUC | F1 | Precision | Recall |
|---|---|---|---|---|
| SVM | 0.86 | 0.86 | 0.86 | 0.86 |
| KNN | 0.77 | 0.80 | 0.71 | 0.91 |
| Decision Tree | 0.74 | 0.75 | 0.71 | 0.81 |
| Random Forest | 0.99 | 0.99 | 1 | 0.98 |
| Naive Bayes | 0.66 | 0.68 | 0.64 | 0.72 |

## V. Conclusion

In this study, we proposed a feature extraction strategy. Our approach was validated on datasets and testing procedures for its applicability and robustness. The main contribution of the paper was the incorporation of machine learning techniques that produced better results. The synthesis of several characteristics enhanced the performance of our method in terms of recognition rate, SVM, KNN, Decision Tree,Random Forest and Naive Bayes. The focus of future study will be on using bio-inspired methods to make integrated features less dimensional. We'll also make an effort to include more tactics that could improve our performance.

## References

[1] W. H. Organization, *Malaria microscopy quality assurance manual-version 2*. World Health Organization, 2016.

[2] Who, "World malaria report 2020: 20 years of global progress and challenges," pp. 1–151, 2020.

[3] H. Gelband, I. I. Bogoch, P. S. Rodriguez, M. Ngai, N. Peer, L. K. Watson, and P. Jha, "Is malaria an important cause of death among adults?" *The American Journal of Tropical Medicine and Hygiene*, vol. 103, no. 1, p. 41, 2020.

[4] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for

detecting malaria," *Translational Research*, vol. 194, pp. 36–55, 2018.

[5] D. Payne, "Use and limitations of light microscopy for diagnosing malaria at the primary health care level," *Bulletin of the World Health Organization*, vol. 66, no. 5, p. 621, 1988.

[6] S. R. Meshnick, T. Taylor, and S. Kamchonwongpaisan, "Artemisinin and the antimalarial endoperoxides: from herbal remedy to targeted chemotherapy," *Microbiological reviews*, vol. 60, no. 2, pp. 301–315, 1996.

[7] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 science and information conference*. IEEE, 2014, pp. 372–378.

[8] T. S. Qaid, H. Mazaar, M. Y. H. Al-Shamri, M. S. Alqahtani, A. A. Raweh, and W. Alakwaa, "Hybrid deep-learning and machine-learning models for predicting covid-19," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[9] K. Roy, S. Sharmin, R. B. M. Mukta, and A. Sen, "Detection of malaria parasite in giemsa blood sample using image processing," *Int J Comput Sci Inf Technol*, vol. 10, pp. 55–65, 2018.

[10] T. F. Scherr, S. Gupta, D. W. Wright, and F. R. Haselton, "Mobile phone imaging and cloud-based analysis for standardized malaria detection and reporting," *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[11] P. Krishnadas, K. Chadaga, N. Sampathila, S. Rao, and S. Prabhu, "Classification of malaria using object detection models," in *Informatics*, vol. 9, no. 4. MDPI, 2022, p. 76.