Open Peer Review on Qeios

# A Graphical User Interface Based on Logistic Regression Approach for Malarial Detection

Manish Kumar[1], Bikash Sarkar[1]

1 Birla Institute of Technology, Mesra

Funding: The authors acknowledge the support of funding agency IDEA: Technology Innovation Hub@ Indian Statistical Institute, Kolkata to continue this work as the part of data science project.

Potential competing interests: No potential competing interests to declare.

## Abstract

Malaria (a mosquito-infected disease) is one of the deadliest communicable diseases in the world. The disease causes a significant global health challenge. According to the World Health Organisation (WHO), millions of deaths occur every year worldwide. The mortality rate poses a challenge to authority and management. Over the years, mathematical and machine learning (ML)-based techniques have been developed to mitigate the scenario. In this study, ML-based prediction techniques are investigated to predict the presence of malaria in individuals. More specifically, three ML-based techniques—Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF)—are employed to differentiate their prediction performance (namely, classification accuracy, precision, recall, and F-score) over a created database (D) consisting of 350 records. Among the adopted techniques, the LR technique shows overall better performance over the test data chosen from D. A graphical user interface (GUI) based on LR is also developed to detect the presence or absence of malaria in any individual. The time spent by the GUI to report the absence or presence of the disease is definitely less than the time spent by malaria experts.

**Manish Kumar**, and **B. K. Sarkar**[*]

*BIT Mesra, India*

[*]Correspondence: bksarkar@bitmesra.ac.in

**Keywords:** Malaria, Communicable disease, Machine Learning, Detection, Graphical Interface.

## 1. Introduction

Malaria poses a significant health challenge worldwide. Billions of people in the world die from this disease annually. Almost 50% of the population in the world is under the threat of Malaria. The disease spreads via infected mosquito bites.

More specifically, the disease is caused by the Plasmodium parasite, which may be of five types, namely, P.falciparum, P.vivax, P.malariae, P.ovale, and P.knowlesi. Among these, P.vivax is most widespread [1][2]. Importantly, the disease commonly occurs in tropical countries. According to WHO, in 2022, there were globally about 249 million cases and 608,000 deaths worldwide [3]. However, malaria is preventable and curable if accurate automated prediction and forecasting models are developed.

Over the years, several mathematical models have been developed to gain insight into disease transmission strategies and to prevent or treat the disease. From the literature, it is found that mathematical modeling of malaria communication started in 1911 with the Susceptible-Infectious-Removed model (SIR), which categorized the hosts and vectors into three groups [4][5]. In recent years, the epidemiology of malaria has been expressed mathematically from unrealistic to high-level models [6]. The important metric called the reproductive number ($R_0$) reports the transmission of a disease. In fact, $R_0$ indicates the expected number of contaminated human hosts after effective mosquito bites in a fully susceptible population [7][8][9]. Thus, $R_0$ furnishes the intensity measure of transmission and reports that an area is disease-endemic areas if its $R_0 >1$ [10]. Interestingly, $R_0$ deduces socioeconomic determinants (such as mortality, mobility, and birth rate) [11].

In order to reduce the affected rate of the disease, effective and timely diagnosis of the disease is essential. Researchers have shown keen interest in developing prediction and forecasting models for malaria using Machine Learning (ML) approaches such as Support vector machine, decision trees, random forest, extreme gradient boosting, logistic regression, k-Nearest Neighbours, Naïve Bayes, and multilayer perceptron based on socioeconomic, climatic and environmental data [12]. Certainly, malaria prediction (or early prediction) models can assist in strengthening prevention and control measures. In 2021, Lee *et al.* [13] investigated machine learning classifiers to predict malaria using patient information from parasite case reports. The study [14] performed an experimental analysis of different machine learning strategies to detect malaria, and the obtained results report that Random Forest is one of the promising learners, offering overall accuracy. Certainly, these strategies have been developed to help improve malaria diagnosis mainly by supporting decision making with respect to microscopic examination. Very recently, Kenia and William [15] propose a deep learning-based approach to detect not only malaria parasites but also leukocytes to perform parasite/µL blood count.

In particular, the existing ML works in the literature are costlier, since most of the works use some image-based tests. With this point in mind, in this paper, ML-based prediction models are first investigated using text data to predict the presence of malaria in individuals. In particular, three ML-based techniques viz. LR, SVM, and RF are applied to compare their prediction performance (namely classification accuracy, precision, recall, and F-score) over a created database (D), consisting of 350 records. A graphical user interface (GUI) based on the logistic regression approach is developed to detect the presence or absence of malaria in any individual. The time spent by the GUI to report the absence or presence of the disease is definitely less than the time spent by malaria experts.

The paper is presented as follows. Section 2 describes the methods and materials, including a description of the dataset, data pre-processing, and the adopted classification algorithms. Section 3 deals with the experimental setup, whereas Section 4 discusses the achieved results and their analysis. Finally, Section 5 concludes the work.

## 2. Materials and Methods

This section describes how data are collected, pre-processed, split for training and test of learners and finally the performance metrices to evaluate the learners for choosing the most appropriate learner for designing GUI. A schematic of the overall work activities of the present work is depicted in Figure 1.

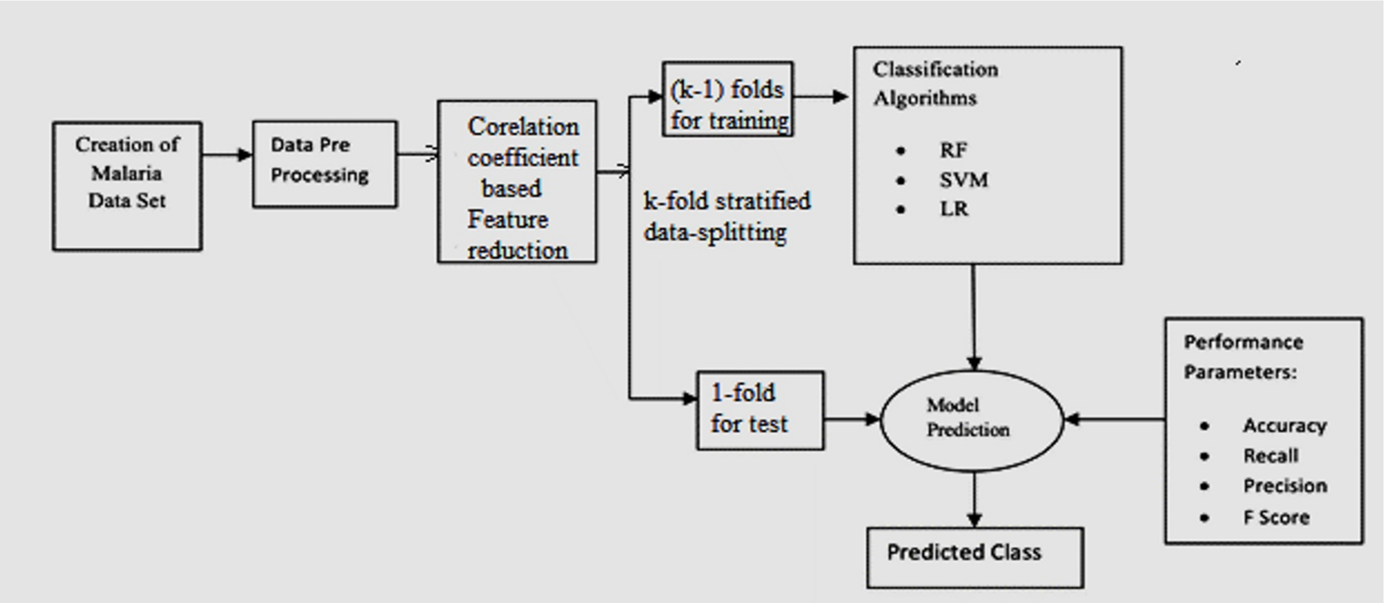


**Figure 1.** A schematic of the overall work activities

Each activity is briefly explained below.

### 2.1. Dataset Description

The present dataset consists of 350 records, of which 178 are male and 172 are female. Out of 350 records, 150 (80 females and 70 males) records are positive (1) malarial cases, and 200 (108 males and 92 females) are negative (0) cases. The data are collected from different health units of Ranchi, Jharkhand, India. The data set initially contains 17 feature attributes with a target class attribute with binary values.

*Data pre-processing technique*: We use the following strategy for handling missing values of attributes.

Missing values of any attribute (A) for the PTB outcome are replaced by the most frequent value of A for the PTB instances; likewise for missing values of any attribute (A) for non-PTB instances. Importantly, the present dataset does not contain missing values and duplicates. But encoding is done in the ‘Gender’ feature.

Further, the present dataset is not a balanced one. That is why a standard stratified k-fold data-splitting technique is used to balance the training set during the learning stage of the learners. In fact, stratified sampling is more functional for operating on imbalanced datasets. It ensures that the training and test datasets possess the same percentage of class labels.

## 2.2. Feature Reduction Technique (FRT)

Feature reduction techniques always improve performance and reduce computational time and overfitting on training data. In particular, FRT discards redundant features that increase the time complexity of machine learning techniques and reduce the overall performance of the learning techniques. The feature reduction technique adopted here is the Pearson correlation coefficient between (non-target, non-target) and (non-target, target) attribute pairs. This is achieved by making a correlation matrix. If the correlation value between any pair (non-target, target) is less than 0.6, then discard the attribute. After applying FTR, we have finally chosen 10 clinical symptoms (features/attribute) which are respectively Hemoglobin, Hematocrit, Mean corpuscular volume(MCV), Platelets, White Blood Corpuscle (WBC), Neutrophils, Lymphocytes, Mean Corpuscular Hemoglobin (MCH), Eosinophils and Basophils for predicting the disease as well designing the GUI.

## 2.3. Classification Algorithms

Three competent ML-based techniques, namely Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF), are here experimented with to compare their prediction performance (namely classification accuracy, precision, recall, and F score) over a created database (D) with 350 persons. The learners are first briefly explained below, and then their achieved performances over the created database are presented in the performance tables.

### 2.3.1. *Support Vector Machine (SVM)*

The SVM learning technique is used both for classification and regression problems. In machine learning, the technique was first developed by Cortes and Vapnik (1995) [16]. The technique is very productive for large-dimensional datasets.

### 2.3.2. *Logistic Regression (LR)*

This statistical model was originally proposed and popularized by Joseph Berkson,[17], in 1944. It is a binary classification task based on statistics, where the goal is to predict the probability of an event based on the features. Though the name implies regression, it is a classification algorithm. The prime benefit of this approach is that it can robustly operate on non-linear data.

### 2.3.3. *Random Forest (RF)*

This technique generates numerous decision trees[18], forming a forest (consisting of m > 1 trees). In fact, it is also known as an ensemble of decision tree algorithms. In particular, individual DTs are highly sensitive to the training data, *i.e.*, bit changes in the training sample, DT structure highly changes. So, our model might fail to generalize. That is why RF is adopted, which is less sensitive to training data. RF applies randomization in two places:

- Random sampling with replacement from the training set is used to construct the trees. It is, indeed, bagging (bootstrap aggregation – sampling with replacement). It reduces sensitiveness to training data.

- While training individual trees, a random subset of features (known as feature randomness; usually*O(logn*) features for each tree, out of *n* total features) is used for searching for splits. This attempts to reduce the association among trees in the random forests (*i.e.*, it avoids a similar act from the trees), which improves the predictive performance.

As different features are used in the trees, the correlation among the trees is reduced. Importantly, by combining these two cases of randomization, RFs are able to reduce overfitting and achieve better performance as compared to individual DTs. In our experiment, 10 decision trees are used in the construction of the forest.

## 3. Experimental Setup

### 3.1. Performance Metrics

The performance of the adopted learners is analysed based on various assessment measures like accuracy, precision, recall, and F1-score and stratified k-fold cross-validation**.** The mathematical expressions of these metrics are expressed in (1)-(4).

$$\text{Classification Accuracy (CA)} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall (or sensitivity or true positive rate: TPR)} = \text{TP}/(\text{TP} + \text{FN}) = \text{TP}/\text{P}$$

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})$$

where TP, FP, TN, and FN denote respectively true positive, false positive, true negative, and false negative.

Importantly, accuracy result reports about the overall power of a learner for predicting the correct class instances, TPR reports about the capability of predicting the target class. Precision provides insights into the model's ability to correctly predict positive instances while minimizing the risk of false alarms.

In k-fold cross validation approach, the original dataset (D) of 350 individuals is split into equal size sub-datasets. The number of subsets depends upon the value of k (*e.g.*, 2, 3, 4,..). Of note, one subset is used for test purposes, and the remaining (k-1) subsets (combined) are used for learning purposes. Each subset (out of k subsets) is distinctly used for test purposes. So, we may capture the mean accuracy for each k-fold. Mean result is taken into account for showing central tendency of the performance, whereas *s.d.* (error) for showing the dispersion of the results around mean. If*s.d.* is very low (*i.e.*, close to zero), it means that the performance results are almost close to each other over different runs. In other words, the achieved results are reliable.

In the present study, values of k are respectively 2, 4, 5, 10. For carrying out the experiments and developing the GUI, Python 3.9 is used.

## 4. Results and Discussion

### 4.1. Results

Based on k-fold (k = 2, 4, 5, 10) cross-validation, the performance measures of the used competent classification algorithms are presented respectively in Tables 1, 2, 3. For each performance metric, a mean value with standard deviation (*s.d.*) is reported.

**Table 1.** Accuracy performance comparison of the classifiers for different k-folds

| K-Fold | SVM<br>(mean-CA% ±s.d.) | LR<br>(mean-CA% ±s.d.) | RF<br>(mean-CA% ±s.d.) |
|---|---|---|---|
| K=2 | 92.28±0.008 | 93.14±0.01 | 90.85±0 |
| K=4 | 91.14±0.01 | 93.70±0.01 | 91.71±0.02 |
| K=5 | 92.28±0.02 | 92.28±0.02 | 91.71±0.02 |
| K=10 | 92.28±0.02 | 93.42±0.02 | 90.28±0.04 |

**Table 2.** Precision performance comparison of the classifiers for different k-folds

| K-Fold | SVM<br>(mean-precision% ±s.d.) | LR<br>(mean-precision% ±s.d.) | RF<br>(mean-precision% ±s.d.) |
|---|---|---|---|
| K=2 | 91.29±0.03 | 92.01±0.01 | 88.37±0.01 |
| K=4 | 90.96±0.05 | 92.63±0.02 | 89.81±0.05 |
| K=5 | 91.45±0.05 | 91.58±0.04 | 91.31±0.07 |
| K=10 | 91.03±0.07 | 91.35±0.04 | 89.11±0.08 |

**Table 3.** Recall performance comparison of the classifiers for different k-folds

| K-Fold | SVM<br>(mean-recall% ±s.d.) | LR<br>(mean-precision% ±s.d.) | RF<br>(mean-precision% ±s.d.) |
|---|---|---|---|
| K=2 | 90.70±0.009 | 92.14±0.03 | 91.34±0.003 |
| K=4 | 88.69±0.03 | 92.78±0.03 | 90.76±0.01 |
| K=5 | 89.85±0.04 | 90.72±0.05 | 87.57±0.03 |
| K=10 | 89.65±0.07 | 93.60±0.03 | 91.44±0.08 |

**Table 4.** F-score performance comparison of the classifiers for different k-folds

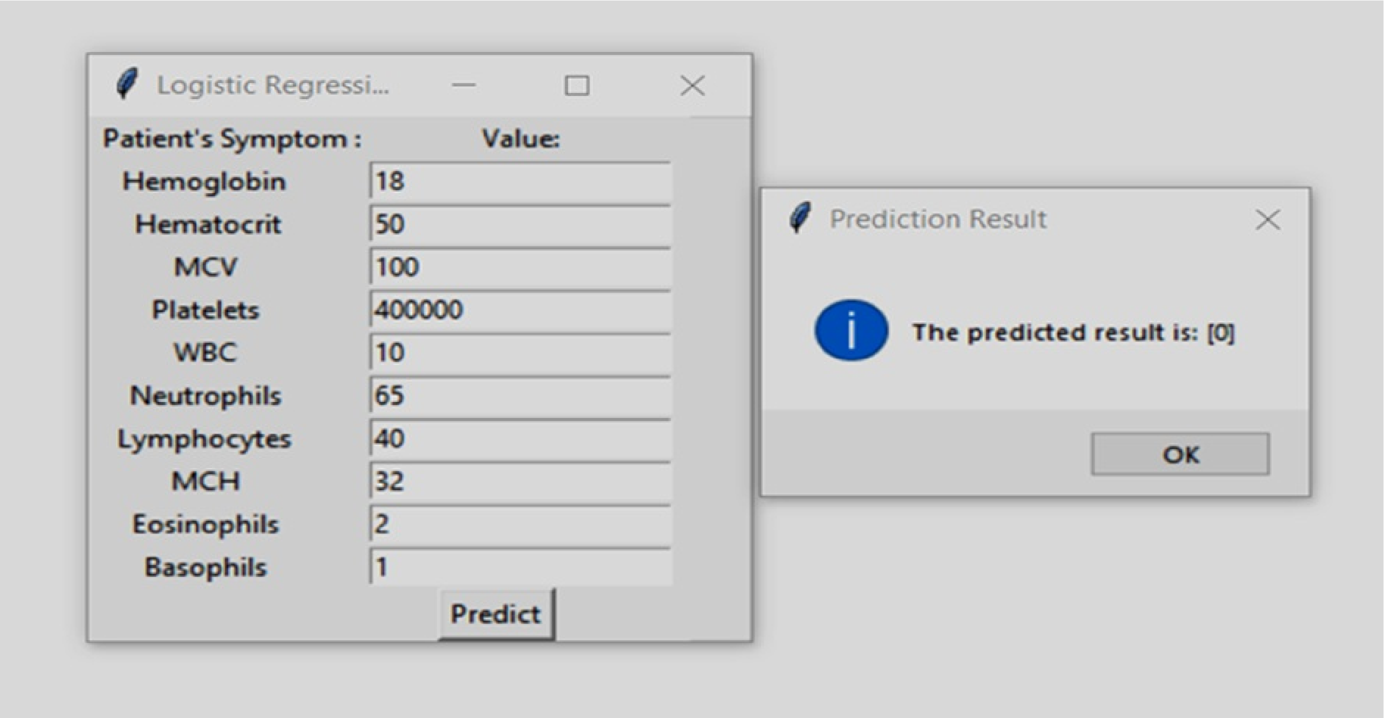
| K-Fold | SVM<br>(mean-F-score% ±s.d.) | LR<br>(mean-F-score% ±s.d.) | RF<br>(mean-F-score% ±s.d.) |
|---|---|---|---|
| K=2 | 90.95±0.01 | 92.01±0.01 | 89.39±0.002 |
| K=4 | 89.56±0.01 | 91.64±0.01 | 90.65±0.02 |
| K=5 | 90.65±0.03 | 90.90±0.02 | 89.05±0.02 |
| K=10 | 90.35±0.04 | 92.34±0.02 | 90.03±0.05 |

### 4.2. Discussion

As presented in Table 1, the maximum*mean* classification accuracy- 93.70% (with very low s.d. 0.01) is yielded by the LR learner. Further, for all k's (2, 4, 5, and 10), it maintains consistency in accuracy. Also, the *s.d.* around CA is very low- this indicates that the classifier LR is much reliable for the prediction of malaria cases.

In Table 2, the precision results for different k-values (2, 4, 5, 10) are presented. The highest mean precision measure is 92.63% (with s.d. 0.02), and it is bagged by the learner logistic regression. Table 3 shows the recall data of the learners for different values of k. The maximum mean recall value is 93.60% (with s.d. 0.03) at k=10, and it is achieved by the LR learner too. Finally, the F-score results of three classification algorithms considering different values of k are presented in Table 4. The maximum mean F-score value, 92.34%, is again yielded by the LR learner.

Now, from all the four tables, it is noticed that logistic regression carries out better amongst all for the experimented values of k (2, 4, 5, and 10). This reveals that LR is the most accurate model among the adopted learners in this study. So, the best performing knowledge yielded by logistic regression (out of k=10 folds) is captured, and a graphical user interface (GUI) based on this knowledge is designed for the prediction of malaria disease, taking less time as compared to the time spent by malaria experts. The GUI is shown in Figure 2.



| Patient's Symptom : | Value: |
|---|---|
| Hemoglobin | 18 |
| Hematocrit | 50 |
| MCV | 100 |
| Platelets | 400000 |
| WBC | 10 |
| Neutrophils | 65 |
| Lymphocytes | 40 |
| MCH | 32 |
| Eosinophils | 2 |
| Basophils | 1 |

The predicted result is: [0]

**Figure 2.** Logistic regression-based GUI for prediction of malaria disease

## 5. Conclusion and Future Scope

The research was undertaken using malaria data, and pre-processing was performed. Further, correlation coefficient analysis was performed for feature selection. On the dataset with selected features, three classification algorithms were applied (SVM, LR, RF) along with k-fold (k=2, 4, 5, and 10) cross-validation. This enabled considerable data analysis to be performed conclusively to determine the optimal result. From the results, it may be concluded that logistic regression performs better in all aspects as compared to other investigated algorithms (in this research) in terms of performance parameters like accuracy, recall, precision, and F1-score. The work claims that the designed GUI will be helpful for medical professionals and healthcare researchers to control malaria disease to a great extent. The time spent by the GUI to report the absence or presence of the disease is definitely less than the time spent by malaria experts.

The data used in this research is very less, since it is collected from only one region. More data may improve the performance of the classifiers. Hence, it will be interesting to use extensive data and data from different areas of future work, and our future work aims to include more data collection, extensive data sampling, and the study with other classifiers for further analysis of the dataset.

## Statements and Declarations

### Acknowledgements

The authors acknowledge the support of funding agency IDEA: Technology Innovation Hub@ Indian Statistical Institute, Kolkata, to continue this work as part of a data science project. The authors acknowledge their parent institute for carrying out the research work.

### Conflicts of Interest

The authors declare no conflict of interest. There is no role of any third party in data collection, analyses, or interpretation of data; in the writing of the manuscript.

## References

1. ^*Antony HA and Parija SC (2016): Antimalarial drug resistance: an overview. Trop Parasitol. 2016;6(1):30. doi: 10.4103/2229-5070.175081.*
2. ^*Jasminka Talapko, Ivana Škrlec, Tamara Alebić, Melita Jukić and Aleksandar Včev (2019): Malaria: The Past and the Present. -Microorganisms. 2019 Jun; 7(6): 179.*
3. ^*https://www.who.int/news-room/fact-sheets/detail/malaria*
4. ^*Ross, R. The Prevention of Malaria; E.P. Dutton & Company: New York, NY, USA, 1910.*
5. ^*Ross, R. Some Quantitative Studies in Epidemiology. Nature 1911, 87, 466-467.*
6. ^*May, R.M. Stability and Complexity in Model Ecosystems; Princeton Landmarks in Biology; Princeton University Press: Princeton, NJ, USA, 2001; ISBN 978-0-691-08861-7.*
7. ^*Aikins, M.K.; Pickering, H.; Greenwood, B.M. Attitudes to Malaria, Traditional Practices and Bednets (Mosquito Nets)*

*as Vector Control Measures: A Comparative Study in Five West African Countries. J. Trop. Med. Hyg. 1994, 97, 81-86.*

8. ^*Smith, D.L.; Hay, S.I.; Noor, A.M.; Snow, R.W. Predicting Changing Malaria Risk after Expanded Insecticide-Treated Net Coverage in Africa. Trends Parasitol. 2009, 25, 511-516.*
9. ^*Smith, D.L.; Dushoff, J.; Snow, R.W.; Hay, S.I. The Entomological Inoculation Rate and Plasmodium Falciparum Infection in African Children. Nature 2005, 438, 492-495.*
10. ^*Smith, D.L.; McKenzie, F.E.; Snow, R.W.; Hay, S.I. Revisiting the Basic Reproductive Number for Malaria and Its Implications for Malaria Control. PLoS Biol. 2007, 5, e42*
11. ^*Agusto, F.B.; Teboh-Ewungkem, M.I.; Gumel, A.B. Mathematical Assessment of the Effect of Traditional Beliefs and Customs on the Transmission Dynamics of the 2014 Ebola Outbreaks. BMC Med. 2015,*
12. ^*Elliot Mbunge, Richard C. Milham, Maureen Nokuthula Sibiya & Sam TakavarashaJr (2023): Machine Learning Techniques for Predicting Malaria: Unpacking Emerging Challenges and Opportunities for Tackling Malaria in Sub-saharan Africa. - Lecture Notes in Networks and Systems (LNNS, volume 724)*
13. ^*You Won Lee a, Jae Woo Choi b c, Eun-Hee Shin (2021): Machine learning model for predicting malaria using clinical information. -Computers in Biology and Medicine, Volume 129,2021. https://doi.org/10.1016/j.compbiomed.2020.104151.*
14. ^*Samir S. Yadav, Vinod J Kadam, Shivajirao M. Jadhav and Sagar Jagtap (2021): Machine Learning based Malaria Prediction using Clinical Findings. -International Conference on Emerging Smart Computing and Informatics (ESCI), AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021.*
15. ^*Kenia H. and William H. (2024): Supporting Malaria Diagnosis Using Deep Learning and Data Augmentation.- Diagnostics,, 14(7), 690;*
16. ^*Cortes, Corinna; Vapnik, Vladimir (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273-297.*
17. ^*Cramer, J. S. (2002): The origins of logistic regression (Technical report). Vol. 119. Tinbergen Institute. pp. 167-178. doi:10.2139/ssrn.360300.*
18. ^*Quinlan, J. R. (1987): Simplifying decision trees. International Journal of Man-Machine Studies. 27 (3): 221-234.*