Comparative Analysis of Malaria Detection Using Predictive Algorithms

Tanay Thatte [1], Ujwal Khairnar [2] & Dr. A.R.Deshpande[3]

1 Tanay Thatte 1; PICT Pune, (Computer Engineering), Pune, Maharashtra , India, tanaythatte17@gmail.com
2 Ujwal Khairnar 2; PICT Pune, (Computer Engineering), Pune, Maharashtra, India, ujwalkhairnar5@gmail.com
3 Dr.A.R. Deshpande 3; PICT Pune, (Computer Engineering), Pune, Maharashtra, India, arativb@gmail.com

Abstract:

Diseases, including Malaria, pose a significant threat to global public health especially in underprivileged communities. They do not affect physical health but also result in economic consequences. This research explores the potential of using data science and machine learning techniques to predict malaria outbreaks. By analyzing datasets containing factors like red blood cells count, white blood cells count , platelet count we developed predictive models to forecast malaria incidents. Models such as Random forest, Gradient boosting, and Support vector machines were tested, and their performance surpassed traditional methods. The findings emphasize the utility of these approaches in proactive public health planning, offering insights for effective resource allocation and intervention strategies.

Keywords: Malaria, Machine Learning, Random Forest, Gradient Boosting, Support Vector Machine.

1 Introduction

1.1 Background

A disease can be defined as any abnormal deviation from the normal functioning of an organism, generally associated with certain symptoms. Malaria is a disease caused by the parasite Plasmodium falciparum . It spread to humans through the bites of infected mosquitoes. Although in most cases Malaria is not very severe, however if proper treatment is not given it can prove to be fatal [10].

Malaria remains a major problem worldwide especially in tropical regions where the weather is humid. Despite many advances in prevention and treatment strategies, the complexity of malaria transmission continues to hinder effective disease control [14].

Machine learning which allows machines to learn and predict values based on past occurrences of a similar event is now being used to predict diseases by feeding past data [9]. This data can range from geographical factors of a particular region where disease occurrence is very high to patient data containing information about their blood tests or X-Ray scan or an MRI scan. By using Machine learning we can identify patterns associated with malaria transmission to enable more effective and targeted prevention interventions of the disease [11].

1.2 Literature Survey

In [1] application of various machine learning algorithms in predicting malaria outbreaks was observed, utilizing factors such as temperature, humidity, and population ratios. Among the algorithms tested, eXtreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), and Random Forests demonstrated the highest predictive capabilities. Evaluation metrics such as accuracy, recall, precision, error rate, and Matthews Correlation Coefficient were used to

comprehensively assess the models. The research emphasizes the significance of data-driven insights to combat the spread of malaria.

In [2] a study was conducted in the Sundargarh district of Odisha, India, where the relationship between climate factors and malaria incidence was analyzed, using WEKA machine learning tools, they compared two techniques, Multi-Layer Perceptron (MLP) and J48, finding that J48 was more effective in predicting malaria with higher accuracy and lower error rates. The study highlighted the significance of seasonal temperature and humidity variations in influencing malaria outbreaks.

In [3] patient data was used to create machine learning models for malaria diagnosis. Race, disease type, gender, age, symptoms were among the patient variables that were identified using information from CDC reports and PubMed abstracts. The performance of six different learning machines—support vector machines, random forests, multilayer perceptrons, gradient boosting, AdaBoost, and CatBoost—is compared in this study. The outcomes show the potential of machine learning in this area by proving the efficiency of random forest models based on patient data for malaria prediction.

In [4] The objective of the study was to predict instances of malaria by making use of machine learning and clinical data. The researchers used various machine learning techniques along with clinical data to build models for predicting malaria. This research emphasizes on the significance of clinical data in predicting malaria by signifying the effectiveness of machine learning in detecting malaria at an early stage. These discoveries portrayed the connection between clinical treatment and disease management technologies, along with the accuracy of machine learning in forecasting malaria cases.

In [5] The 2020 paper authored by Mehmood, Mahmoud, and Adeel, and published in IEEE Access, delves into the realm of machine learning algorithms in the context of predicting malaria epidemics and conducting data analysis. The authors thoroughly explore various methodologies and strategies for forecasting malaria cases, while also scrutinizing pertinent data. Their research revolves around the utilization of machine learning techniques to enhance the precision of malaria prediction models through effective data processing and interpretation.

In [6] The Acta Tropica publication in 2018, authored by Karunamoorthi and Almadiy, provided a comprehensive analysis of contemporary techniques and obstacles in the field of malaria outbreak prediction modeling. The researchers delved into diverse methodologies and strategies employed for forecasting and averting malaria outbreaks, thereby illuminating the prevailing challenges and constraints.

In [7] the paper by Amara and Pradhan provides a systematic review of machine learning techniques applied to malaria risk modeling. This study assessed various machine learning methodologies, such as decision trees, random forests, support vector machines, and others, in the context of predicting and assessing the risk of malaria transmission.

2 Proposed Methods

2.1 Models Used

We have used Predictive Models such as Random Forest, Gradient Boosting and Support Vector Machine as these models are able to analyze large datasets, identify risk factors, and forecast the likelihood of disease occurrence. By integrating various clinical, genetic, and lifestyle data, these models can provide personalized risk assessments and early warnings [12, 13].

2.1.1 Random Forest

An ensemble learning approach called Random Forest bootstraps random subsets of the training data and takes random feature subsets into consideration at each split to create numerous decision trees [8]. Because the trees are more diverse due to this unpredictability, overfitting is less likely. Each tree "votes" for a class in classification tasks, and the final prediction is the majority class; in regression tasks, predictions are averaged. Combining several trees improves the model's robustness and accuracy, which makes Random Forests useful for a range of applications while reducing the drawbacks of single decision trees.

Step By Step Working :

1. **Data Bootstrapping:**
    - Randomly sample the training dataset with replacement, creating multiple bootstrap samples.
2. **Random Feature Selection:**
    - For each bootstrap sample, randomly select a subset of features at each split when building a decision tree.
3. **Decision Tree Building:**
    - Build a decision tree for each bootstrap sample using the selected features. Grow the tree until a certain criterion is met (e.g., maximum depth).
4. **Ensemble of Trees:**
    - Create an ensemble of decision trees by repeating steps 1-3, resulting in multiple diverse trees.
5. **Voting (Classification) or Averaging (Regression):**
    - For classification tasks, let each tree "vote" for a class, and the majority class becomes the final prediction. For regression tasks, average the predicted values from all trees.
6. **Aggregation:**
    - Combine the predictions of all trees to obtain the final prediction, providing a robust and accurate model.

$$F(X) = (\sum_{i=1}^{N} F_i\,(X))/(N) \quad \textbf{(1)}$$

Random Forest

X represents the input

$F_i(X)$ is the prediction of the i^{th} decision tree

N is the total number of decision trees in the Random Forest

2.1.2 Gradient Boosting

Gradient boosting is a powerful ensemble technique that corrects the errors of its predecessors by constructing a sequence of weak learners, typically decision trees, one by one. Gradient boosting involves fitting each new tree to its residuals, which are the differences between actual and expected values. The goal of each new tree in the gradient boosting iterative process is to minimize the errors caused by the collection of previous trees. Gradient boosting achieves high predictive precision and robustness by combining weak models in a weighted combination, where each tree improves the overall prediction. However, it tends to overfit if not regularized, and requires careful hyperparameter tuning. The efficiency and scalability of the gradient boosting algorithm have been greatly improved by well known implementations such as XGBoost or LightGBM.

Step By Step Working :

1. **Initialize the Model:**
    - Start with a simple model, usually a constant value (mean for regression problems or a class with the highest frequency for classification).
2. **Compute Residuals:**
    - Calculate the residuals by subtracting the predicted values from the actual target values.
3. **Build a Weak Learner:**
    - Train a weak learner (typically a shallow decision tree) on the residuals. The goal is to fit the model to the errors made by the current model.
4. **Compute the Learning Rate Multiplier:**
    - Introduce a learning rate (η), a small positive number less than 1, to control the step size in updating the model. This helps prevent overfitting and stabilize the learning process.
5. **Update the Model:**
    - Update the current model by adding the learning rate multiplied by the predictions of the weak learner to the previous model's predictions. This step minimizes the residuals.
6. **Repeat Steps 2-5:**
    - Repeat steps 2-5 until a specified number of weak learners are trained or until a certain criterion is met (e.g., achieving satisfactory performance).
7. **Final Prediction:**
    - The final prediction is the sum of the predictions from all the weak learners. For regression tasks, it's a continuous value, and for classification tasks, it's converted into probabilities or class labels.

$$F(X) = \sum_{i=1}^{N} \eta . f_i(X) \quad \textbf{(2)}$$

Gradient Boosting

F(X) is the final prediction

$f_i(X)$ represents the prediction of the i^{th} weak learner

η represents the learning rate, a small positive value that scaler the value of each learner

2.1.3 Support Vector Machine

Support vector machines (SVM) are supervised machine learning algorithms that attempt to find the best hyperplane in the space of data to distinguish different classes of data by maximizing the margin, which is the distance from the hyperplane to the nearest data point from each class, influenced by support vectors (the closest data points).SVM can work with non-linear separable data by transforming the feature space using a kernel function. The algorithm works well in high-dimensional space and can be used for classifying and regression tasks.SVM is versatile, but its performance is dependent on the kernel and parameter choices and it may not work well on large datasets.Despite these drawbacks, SVM is still widely used in various domains because of its robustness and its generalization capabilities.

Step By Step Working:

1. **Data Representation:**
    - Represent each data point as a feature vector in a multidimensional space.
2. **Initialization:**
    - Choose an initial hyperplane that separates the classes.
3. **Margin Maximization:**
    - Identify the support vectors (closest points to the hyperplane) and maximize the margin between classes.
4. **Optimization:**
    - Formulate an optimization problem to find the optimal hyperplane weights that maximize the margin while minimizing errors.
5. **Optimization Solving:**
    - Solve the optimization problem to obtain optimal weights and biases.
6. **Final Hyperplane:**
    - The optimal hyperplane is determined by the obtained weights, maximizing separation.
7. **Decision Function:**
    - Define a decision function based on weights and biases.

8. **Classification:**

- Classify new data points based on the sign of the decision function.

2.2 Dataset Used

For this research we have used hematological data collected from Ghana. Hematological Data tells us information about the blood samples collected.

2.2.1 Size

The dataset used has 2207 rows and 34 columns.

2.2.2 Parameters

The parameters present in our dataset are :

'SampleID', 'consent_given', 'location', 'Enrollment_Year', 'bednet',

'fever_symptom', 'temperature', 'Suspected_Organism',

'Suspected_infection', 'RDT', 'Blood_culture', 'Urine_culture',

'Taq_man_PCR', 'parasite_density', 'Microscopy', 'Laboratory_Results',

'Clinical_Diagnosis', 'wbc_count', 'rbc_count', 'hb_level',

'hematocrit', 'mean_cell_volume', 'mean_corp_hb', 'mean_cell_hb_conc',

'platelet_count', 'platelet_distr_width', 'mean_platelet_vl',

'neutrophils_percent', 'lymphocytes_percent', 'mixed_cells_percent',

'neutrophils_count', 'lymphocytes_count', 'mixed_cells_count',

'RBC_dist_width_Percent'

The output will be given by the Clinical Diagnosis column. As input to our models we will use the parameters :

'wbc_count', 'rbc_count', 'hb_level',

'hematocrit', 'mean_cell_volume', 'mean_corp_hb', 'mean_cell_hb_conc',

'platelet_count', 'platelet_distr_width', 'mean_platelet_vl',

'neutrophils_percent', 'lymphocytes_percent', 'mixed_cells_percent',

'neutrophils_count', 'lymphocytes_count', 'mixed_cells_count',

'RBC_dist_width_Percent'

All these parameters can be directly obtained from a blood sample [15].

2.3 Preprocessing Techniques

From our dataset our output value will be 'Clinical Diagnosis'. We will consider parameters such as white blood cells count, red blood cells count, hemoglobin level, mean cell volume and all the values which can be derived from a blood test. We will use these parameters to train our model. For better accuracy we will convert our data into an integer value and scale it between 0 and 1. We will then divide our dataset into training dataset and testing dataset in a ratio of 80:20.

3. Results & Discussion

We have passed our training dataset into our Models which are Random Forest, Gradient Boosting and Support Vector Machine. We then test the accuracy of our model by comparing it with our testing dataset. There are 3 possible outputs which are :

- Severe Malaria – When malaria is present and very severe such that it can lead to death of the patient
- Uncomplicated Malaria – When malaria is present but not severe
- Non-Malaria Infection – When malaria is not present in the patient

3.1 Random Forest

Table 1 Random Forest Results

Type	Precision	Recall	F1 Score
Severe Malaria	0.94	0.97	0.96
Uncomplicated Malaria	0.81	0.66	0.73
Non-Malaria Infection	0.81	0.90	0.86

On testing with Random Forest Model we get the above results. The overall Accuracy of Random Forest model is 0.84.

3.2 Gradient Boosting

Table 2 Gradient Boosting Results

Type	Precision	Recall	F1 Score
Severe Malaria	0.96	0.95	0.95
Uncomplicated Malaria	0.73	0.66	0.69
Non-Malaria Infection	0.80	0.85	0.83

On testing with Gradient Boosting Model we get the above results. The overall Accuracy of Gradient Boosting model is 0.83.

3.3 Support Vector Machine

Table 3 Support Vector Results

Type	Precision	Recall	F1 Score
Severe Malaria	0.94	0.95	0.94
Uncomplicated Malaria	0.79	0.66	0.71
Non-Malaria Infection	0.80	0.89	0.85

On testing with Support Vector Machine Model we get the above results. The overall accuracy of Support Vector Machine Model is 0.83.

3.4 Testing with variation of Training and Testing Datasets

We have also tested how the accuracy of a model changes when we change the ratio of training size : testing datasize. The results are as follows :

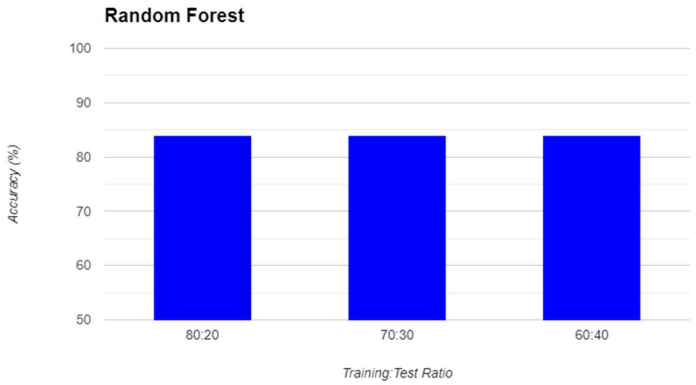

Fig 1 Random Forest Testing

Fig 2 Gradient Boosting Testing

Even after altering the training : testing ratios Random Forest is the most efficient model. For all 3 models we can see that we get the highest accuracy when the Training : Testing ratio is 70 : 30.

3.5 Discussion

We have tested the accuracy with same training and testing dataset for all 3 of our models. On comparing from Table 1, Table 2 and Table 3 we can see that Random Forest model gives the highest accuracy of 84% compared to Gradient Boosting's 81% and Support Vector Machine's 83%.

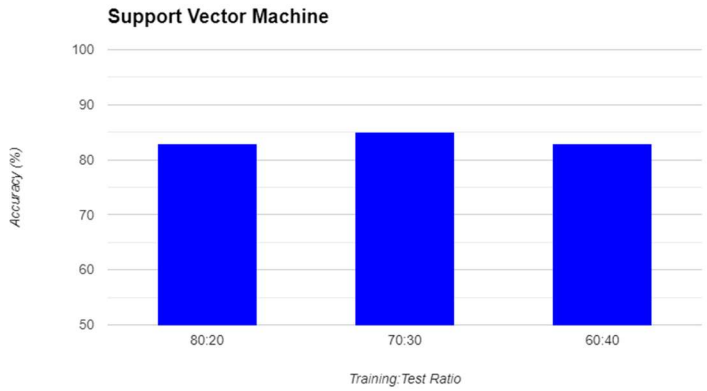

Fig 3 Support Vector Machine Testing

We also decided to change the training: testing ratios to see the variation of accuracy in the models. On changing the ratio to 70: 30 we can observe from Fig 1, Fig 2 and Fig 3 that accuracy of Random Forest remains constant at 84%, whereas accuracy of Gradient Boosting and Support Vector Machine increase to 85%. When the ratio was made to 60: 40 we observed that the accuracy of Random Forest remained constant at 84%, whereas accuracy of both Gradient Boosting and Support Vector Machine dropped to 83%.

We can say that Random Forest Model provides the most accurate predictions and it's accuracy does not vary on changing the training: testing ratios.

4 Conclusion

In conclusion our research demonstrates the use of machine learning in predicting malaria outbreaks. We have tried to predict malaria by passing data related to patient's blood.

For a Machine Learning Model an accuracy of 84% might be good, however when dealing with a critical issue such as disease prediction we should always aim at improving our model for a better accuracy as it can help in saving lives. In future we can try on expanding datasets and using more diverse datasets.

References

[1] Godson Kalipe, Vikas Gautham, Rajat Behera "Predicting Malaria outbreaks Using Machine Learning and Deep Learning." in IEEE 2018 International Conference on Information Technology (ICIT) 2018.

[2] Pallavi Mohapatra, Nitin Kumar Tripathi, Indrajit Pal, Sangam Shrestha "Comparative Analysis of Machine Learning Classifiers for the Prediction of Malaria Incidence Attributed to Climatic Factors." Research Square 2020.

[3] You Won Lee , Jae Woo Choi , Eun-Hee Shin "Machine learning model for predicting malaria using clinical information." ScienceDirect 2020.

[4] Samir S. Yadav, Vinod J Kadam, Shivajirao M. Jadhav, Sagar Jagtap, Prasad R. Pathak "Machine Learning based Malaria Prediction using Clinical Findings." in IEEE 2021 International Conference on Emerging Smart Computing and Informatics (ESCI)

[5] Mehmood, I., Mahmoud, M. S., & Adeel "Malaria prediction and data analysis using machine learning techniques." IEEE Access, 8, 124223-124241.

[6] Karunamoorthi, K., & Almadiy, A "Malaria outbreak prediction modeling: an overview of the recent approaches and challenges." Acta Tropica 2018.

[7] Kah Yee Tai & Jasbir Dhaliwal "Machine learning model for malaria risk prediction based on mutation location of large-scale genetic variation data." Springer 2022.

[8] MDPI "Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images": https://www.mdpi.com/2072-4292/14/3/574, Jan.25 2021 [Nov.1 2023]

[9] Thakur S, Dharavath R "Artificial neural network based prediction of malaria abundances using big data: a knowledge capturing approach." Clin Epidemiol Glob Health. 2019.

[10] Sharma V, Kumar A, Panat L, Karajkhede G, Lele A. "Malaria outbreak prediction model using machine learning." Int J Adv Res Comput Eng Technol. 2021.

[11] Poostchi M, Silamut K, Maude RJ, Jaeger S "Image analysis and machine learning for detecting malaria." Transl Res. 2018.

[12] Adebiyi MO, Arowolo MO, Olugbara O "A genetic algorithm for prediction of RNA-seq malaria vector gene expression data classification using SVM kernels." Bull Electr Eng Inform. 2021.

[13] Wojciech Siłka, Michał Wieczorek, Jakub Siłka andMarcin Woźniak "Malaria Detection Using Advanced Deep Learning Architecture" MDPI 2023.

[14] Manjurano A, Clark TG, Nadjm B, Mtove G, Wangai H, Sepulveda N "Candidate human genetic polymorphisms and severe malaria in a Tanzanian population" PLOS ONE 2012.

[15] Manas Kotepui, Duangjai Piwkham, Bhukdee PhunPhuech, Nuoil Phiwklam, Chaowanee Chupeerach and Suwit Duangmano "Effects of Malaria Parasite Density on Blood Cell Parameters" PLOS ONE 2015.