Scientific Research Publishing

Open Journal of Applied Sciences, 2025, 15(6), 1752-1765
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

Comparing Classification Models for Predicting Malaria: A Case Study of Malaria Incidence in Kenya

Shalyne G. Nyambura[1], Kinya Kaibung'a[2], Annette N. Nyambura[3]

[1]Department of Mathematics, Meru University of Science & Technology (MUST), Meru, Kenya
[2]Department of Mathematics, Pan African University Institute of Basic Sciences Technology and Innovations (PAUSTI), Nairobi, Kenya
[3]Department of Business and Economics, Meru University of Science & Technology (MUST), Meru, Kenya
Email: nshalyne@gmail.com

How to cite this paper: Nyambura, S.G., Kaibung'a, K. and Nyambura, A.N. (2025) Comparing Classification Models for Predicting Malaria: A Case Study of Malaria Incidence in Kenya. *Open Journal of Applied Sciences*, **15**, 1752-1765.
https://doi.org/10.4236/ojapps.2025.156120

Received: May 16, 2025
Accepted: June 24, 2025
Published: June 27, 2025

Abstract

Accurate prediction of malaria incidence is indispensable in helping policy makers and decision makers intervene before the onset of an outbreak and potentially save lives. Various classification models have been fitted in prior studies to describe both climatic and non-climatic factors that influence the spread of malaria. However, there have been no comprehensive studies comparing classifier algorithms to find which best describes the spread of malaria. This study fits five machine learning models to real malaria data for Kenya in a bid to establish which model has the highest predictive capability in the presence of various climatic and non-climatic predictor variables. Methods described in the available literature to forecast malaria involve complex numerical simulations for malaria transmission. This study presents a data-driven binary classification approach for the prediction of malaria incidence. Various metrics based on the confusion matrix are computed and used as a basis for comparing the fitted models. The results infer that Random Forest is the best model with 76% accuracy, 64.67% precision, 40.75% recall, and 90.71% specificity.

Keywords

Machine Learning, Binary Classification, Random Forest, Malaria Prediction, Confusion Matrix

1. Introduction

Malaria is a life-threatening parasitic infection common in tropical countries that

is transmitted to humans through bites of female *Anopheles mosquitoes* infected with the plasmodium parasite. According to the World Health Organization, there were approximately 229 million cases of malaria worldwide in 2019 while deaths stood at 409,000 [1]. Children aged under 5 years are the most vulnerable group, accounting for 67% (274,000 cases) of all malaria deaths worldwide in 2019. Closer home, malaria is a major cause of morbidity and mortality in Kenya with more than 70 percent at risk of contracting the disease and, mostly around Lake Victoria and coastal regions [2]. According to the Ministry of Health Kenya (Malaria Control Programme), more than four million cases of malaria are reported annually in Kenya. A mortality rate of 5.1% has been reported among patients admitted with severe malaria in Kenyan hospitals.

The Kenya Malaria Strategy (2019-2023) was implemented to reduce the incidence rates and fatalities of malaria by 75% by the close of 2023 [3]. Despite these efforts, malaria remains a major issue in the country, with roughly 5.5 million cases reported in 2023 alone. The government is combating these cases through stringent measures such as availing insecticide-treated nets, indoor residual spraying, robust case management, and seasonal malaria chemoprevention. The President's Malaria Initiative, backed by the US government, supported spraying efforts in Migori and Busia counties [4].

Albeit life-threatening, malaria is both preventable and curable. However, the spread of malaria is hard to curb due to the communicable nature of the disease. As such malaria incidence should be treated as a potential medical emergency with prompt treatment to avoid further spread of the disease. Malaria can be suspected based on a patient's travel history, the physical findings at examination and symptoms. However, for a definitive diagnosis to be made, laboratory tests must be carried out to demonstrate the malaria parasites or their components. Besides the presence of plasmodium parasites, lifestyle and environmental factors, including altitude, rainfall, and biting intensity, influence the transmission pattern.

Malaria incidence in Kenya is widespread in areas around the coast and Lake Victoria, bearing the greatest burden of the disease. The Ministry of Health has implemented several sound policies and evidence-based strategies in the fight against malaria through the National Malaria Control Programme. However, these efforts have been marred by various challenges, such as parasite resistance to drugs, lack of effective and lasting prevention strategies, resistance to insecticides, and the impact of climate change. Successful mitigation of the spread of malaria is contingent on efficient modeling of the transmission patterns while factoring in the climatic and non-climatic factors.

1.1. Factors Influencing Malaria Incidence

Several research scholars across the world have revealed numerous factors that are associated with the factors influencing malaria.

1.1.1. Climatic Factors

1) Temperature

The time required for the parasite to complete its development in the mosquito's gut is about ten days, but it can be shorter or longer than that, depending on the temperature [5]. As the temperature decreases, the number of days necessary to complete the development increases for a given Plasmodium species. P. Vivax and *P. falciparum* have the shortest development cycles and are, therefore more common than P. Ovale and P. Malariae. Malaria transmission in colder 8 areas can sometimes occur because the Anopheles often live in houses that tend to be warmer than the outside temperature. Higher temperatures also increase the number of blood meals taken and the number of eggs laid by the mosquitoes, which increases the number of mosquitoes in each area [6].

2) Altitude

Altitude influences the distribution and transmission of malaria indirectly through its effect on temperature [7]. As the altitude increases, the temperature decreases, so the highlands become colder and the lowlands warmer. For instance, in Ethiopian highlands, with altitudes between 2000 and 2400 meters, malaria transmission occurs for short periods only when temperatures rise unusually high [8]. In Kenya, the Great Rift Valley is another location with the perfect altitude for Anopheles mosquito to flourish. This geological trench has lakes, volcanoes, and hot springs spanning several counties. Data shows these regions have high incidences of malaria transmission [9].

3) Rainfall

Studies show that *Anopheles mosquitoes* breed in water. Different *Anopheles mosquitoes* prefer different types of water bodies in which to breed. So, the right amount of rainfall is often essential for them to breed. There are also places where less rainfall and drought can favor mosquito breeding and malaria transmission [10]. Such places are usually covered by vegetation throughout the year, and streams and rivers often flow rapidly. A muddy rainwater collection can support mosquito breeding if it is not polluted.

4) Relative humidity

When humidity is 0% (low), this means that the air is completely free from moisture whereas when humidity is 100% (high) it means that the air is completely saturated with moisture. Mosquitoes survive better under conditions of high humidity [10]. This is why they are more active and prefer feeding during the night.

1.1.2. Non-Climatic Factors

1) Malaria Vectors

Different species of *Anopheles mosquitoes* differ in their capacity to transmit malaria. Mosquitoes in the *Anopheles gambiae* group are the most efficient malaria vectors in the world [11]. The higher incidence of malaria in Africa compared to other parts of the world is mainly due to the efficacy of the mosquitoes transmitting the parasites. Mosquitoes that mainly feed on humans are more efficient carriers of malaria than those that feed on animals. Therefore, the type of *Anopheles mosquitoes* influences the intensity of malaria transmission. Malaria Para-

sites: Studies show that there are four types of malaria parasite that can infect people. These parasites are single-celled. One species—*P. falciparum*—is more common in Africa than in other parts of the world [12].

2) Human Migration

Population movements have greatly influenced malaria transmission. Poor living conditions and inadequate health care worsen the burden of malaria. Migrants from malaria-free highlands have a low immunity against malaria. In addition, major environmental activities such as deforestation enhance the proliferation of mosquito breeding sites thus causing malaria outbreaks. Displaced people from infected areas can reintroduce malaria into areas that are malaria free. According to [13] human migration can worsen the problem of malaria.

3) Project Developments

Water-related development projects such as irrigation channels, dams and ponds can increase the incidence of malaria. Irrigation facilitates breeding sites for malaria mosquitoes leading to increased malaria transmission [14]. For instance, the use of irrigation to flood agricultural land during rice cultivation has long been associated with an increase in the number of vectors and a corresponding increase in the burden of malaria.

4) Drug resistance in malaria parasites

Parasites can also develop resistance drugs or to medicines that are designed to harm them. As a result, the parasites inside the human body can no longer be harmed and patients cannot be cured unless new drugs are developed for treatments. If drug-resistant malaria parasites are not cleared by treatment from infected individuals, they are easily picked up by vector mosquitoes and transmitted to new susceptible individuals who then develop drug-resistant malaria [15].

5) Human host factors

When it comes to malaria, people are either immune, or non-immune. Immune people often have a better chance of tolerating the effects of malaria and surviving the disease than non-immune people. In highly endemic areas, children under five years of age and pregnant women are the most at risk since they have weak immunity to malaria infection [16]. Immunity against malaria develops slowly after several infections and children need at least five years to develop their immunity. Pregnant women have less immunity against malaria. The government is conducting molecular surveillance of *Anopheles mosquitoes* and devising solutions to protect expectant women [17].

2. Methodology

Classification models are a type of machine learning algorithms that are used to predict the category or class that a given input belongs to. In classification, the output variable (or target) is a discrete label or class. The goal of a classification model is to learn the relationship between input features and the target class, so that it can predict the class for new, unseen data. This study applies supervised learning approaches where the model is trained on labeled data (*i.e.*, data that in-

cludes both features and the corresponding class labels). Once trained, the model can make predictions on new, unseen data by assigning a class label to each new input. For each model fitted, a confusion matrix is developed and various metrics including accuracy, recall, specificity and precision are computed to determine which classifier is best suited for prediction of malaria.

2.1. Classification Models Considered

This study considers five widely used supervised classification models to cover both linear and non-linear decision boundaries as well as both instance-based and ensemble approaches, namely logistic regression, Decision trees, Random forest, Support vector machine, and Naïve bayes model as described in [18] and [19]. Other potential models initially considered were XGBoost and multilayer neural networks. However, we deferred these to future work to keep the workflow transparent and interpretable for public-health stakeholders in the current application. A brief description of each classifier is given here:

- Logistic Regression: A baseline linear statistical model used for binary classification problems, where the outcome is a probability that can be mapped to a class label (usually 0 or 1).
- Decision Trees: A tree-like model that splits the data into branches based on feature values, ultimately assigning a class to each leaf node.
- Random Forests: An ensemble tree method that combines multiple decision trees to improve classification accuracy by averaging or voting on their outputs. It is best known for its strong out-of-sample performance.
- Support Vector Machines (SVM): A max-margin classifier that finds the hyperplane that best separates classes in a high-dimensional feature space.
- Naive Bayes: A probabilistic model based on Bayes' theorem, often used for text classification tasks like spam detection.

2.2. Confusion Matrix

A confusion matrix is a table often used to describe the performance of a classifier on a set of test data for which the true values are known. In this study confusion matrices were obtained for binary classifiers. Regarding malaria incidence, the classifier may label the subject status as: *Yes—subject has malaria* or *No—Subject does not have malaria*, while the known/true status of the individual may be Infected/Not infected.

Actual/known status	**Predicted status**	
	Yes (Has malaria)	No (Does not have malaria)
Infected (Actual positive)	True Positive (TP)	False Negative (FN)
Not infected (Actual negative)	False Positive (FP)	True Negative (TN)

The matrix shows the number of true positives, false positives, true negatives, and false negatives, where:

- True positive (TP) refers to cases in which the classifier predicted yes and the individual is in real sense infected with malaria.
- True negative (TN) refers to cases in which the classifier predicted no and the individual in real sense is not infected with malaria.
- False positive (FP) refers to cases in which the classifier predicted yes but in real sense the individual does not have malaria (Type I error).
- False negative (FN) refers to cases in which the classifier predicted no but in real sense the individual has malaria (Type II error).

The confusion matrix helps to identify model errors and specify the types of errors (false positives or false negatives) are more prone to. This can be especially important in applications where one type of error is more costly or dangerous than another (e.g. diagnosing a disease where false negatives might be more critical than false positives).

2.3. Model Evaluation Metrics

To assess how well each binary classification model performs, several metrics are computed based on the confusion matrices obtained, as discussed in [20] and [21].

- **Accuracy**: The proportion of correctly predicted instances over all instances. This metric shows, generally, how often the classifier is correct in its prediction.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \tag{1}$$

- **Precision (Positive Predictive Value)**: The proportion of true positives among all positive predictions. This metric shows how often the model is correct when it predicts yes or how accurate your model's positive predictions are.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

- **Recall (Sensitivity/True Positive Rate)**: The proportion of true positives among all actual positives. This metric shows how many of the actual positive cases were correctly identified using the model or how good your model is at identifying all the positive cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

- **F1-Score:** This metric provides a single score, based on the harmonic mean, that balances both precision and recall, especially where there is an imbalance between classes. It is particularly useful when you want to consider both the false positives and false negatives in a balanced way and when n**either precision nor recall alone is sufficient** to describe the model's performance. The **harmonic mean** tends to be lower than the arithmetic mean, especially when precision and recall differ significantly. This penalizes situations where one of the metrics (precision or recall) is much worse than the other. The harmonic mean ensures that a model can only achieve a high F1-Score if it performs well in both precision and recall.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

- **Specificity/true negative rate:** This is the proportion of actual negative cases correctly identified by the model. This metric refers to the proportion of individuals in the sample data who received a negative result on the test out of those who do not actually have malaria. When the individual's actual disease status is not infected, how often does the model predict no? TNR is equivalent to 1 minus False Positive Rate

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

- **Misclassification Rate:** Generally, how often is the classifier wrong? Equivalent to 1 minus accuracy

$$\text{Misclassification Rate} = \frac{\text{FP} + \text{FN}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \tag{6}$$

- **Kappa coefficient:** This metric gives the measure of agreement between variables and can be computed as follows:

$$K = \frac{\text{P agree} - \text{P chance}}{1 - \text{P chance}} \tag{7}$$

A Kappa value close to 1 indicates a strong agreement between predicted and actual values, adjusting for chance. A Kappa coefficient of zero in this case represents agreement equivalent to chance between the climatic and non-climatic variables.

3. Results and Discussion

This section presents various outcomes from the data analysis to predict whether an individual has or does not have malaria. Secondary data was obtained from the Kenya Malaria Indicator Survey, 2015 and analyzed using R statistical software. Five classification models discussed in section 2 were fitted to sample data on malaria status of 2097 individuals along with other characteristic features. The data was subdivided into training data (70 percent) and testing data (30 percent). There were no missing values in the dataset. We used a stratified 10-fold cross-validation framework to assess out-of-sample performance for each of the five classifiers considered. The full dataset was first shuffled (taking a random seed number as 42) and split into 10 equally sized folds preserving the ration of positive to negative malaria presence indicator. In each iteration, nine folds were used to train the model while the remaining fold was used to test the model. The performance metrics such as accuracy, sensitivity, specificity, and F1-score were averaged over all folds to ensure robust estimates of generalization. Rather than using default parameters, all five classifiers were optimized via a grid search within each training fold, using a nested (inner) 5-fold CV to select parameters minimizing log-loss (for probability models) or maximizing accuracy (for non-probabilistic ones). For the probability-based classifiers (Logistic Regression, Random Forest, and SVM with probability outputs), we varied thresholds from 0.1 to 0.9, observing the per-

formance measures. We found that the default 0.5 cutoff balanced precision and recall best, while the F_1 score was maximized at the 0.48 threshold, though only a difference of less than 0.2% was observed compared to when the threshold was set at 0.5. All reported metrics use threshold = 0.5 for comparability.

In the Kenya MIS 2015 sample, there was an imbalance in the classes with malaria-infected (positives) comprising approximately 18% of observations while the remaining 82% were malaria-free (negatives). We explicitly accounted for this imbalance in two ways: For tree-based methods (Random Forest, Decision Tree), we set the class weight equal to "balanced" so that the minority class received proportionally higher weight during split-criterion calculation. For Logistic Regression and SVM, we likewise set the class weight equal to "balanced" as well. Additionally, we experimented with SMOTE oversampling on the training folds, but performance gains were marginal (<0.5% in F_1), so our primary results use class-weighting only.

3.1. Study Variables

One dependent variable and ten independent variables, inclusive of both climatic and non-climatic factors, were considered in the model fitting. Table 1 provides a summary of the variable types and levels where applicable. All categorical predictors including wealth index, malaria endemicity, education level, and mosquito net ownership were one-hot encoded. Binary indicator columns were created for each level of the categorical variables, dropping the reference level to avoid multicollinearity. Numeric variables such as altitude were normalized to transform into 0 - 1 continuous variables.

Table 1. Description of study variables.

Variable name	Variable type	Description
Age	Discrete	Ranges from 15 - 49 years
Mosquito nets	Categorical	4 levels: No net, only treated nets, both treated and untreated nets, only untreated nets
Szone (Malaria endemicity)	Categorical	Grouped by climatic zone 5 levels: 1-Low risk, 2-Coast endemic, 3-Semi arid, 4-Lake region, 5-Highland
Wealth index	Categorical	Measure of individual wealth level 5 levels: Very poor, Poor, Middle-class, Rich, Very rich.
Education level	Categorical	Highest education attainment. 4 levels: No education, Primary, Secondary, and Higher education.
Cluster altitude	Discrete (Continuous)	Height above sea level Initially measured in meters but normalized to fit in the range of 0 - 1.
Presence of *P. Falciparum*	Categorical	2 levels: 1-Yes or 0-No
Presence of P. Malariae	Categorical	2 levels: 1-Yes or 0-No
Presence of P. Vivax	Categorical	2 levels: 1-Yes or 0-No
Presence of P. Ovale	Categorical	2 levels: 1-Yes or 0-No
Malaria	Dichotomous (dependent variable)	2 levels: 1-Yes/Has malaria or 0-No/Does not have malaria

3.2. Predictive Performance of the Fitted Models

Performance metrics including accuracy, recall, specificity and precision were computed based on the confusion matrices and the results are discussed here for each classifier. Table 2 gives a summary of the performance measures under different models with the comparative abilities displayed in Figure 1.

Table 2. Results of the performance metrics for the five classifiers.

MODEL	ACCURACY	PRECISION/ PPV	RECALL SENSITIVITY	SPECIFICITY
LOGISTIC MODEL	0.7002	0.06038	0.4571	0.7121
SUPPORT VECTOR MACHINE	0.7056	0.0000	0.0000	1.0000
DECISION TREE	0.6989	0.4565	0.1519	0.9276
RANDOM FOREST	0.7600	0.6467	0.4075	0.9071
NAÏVE BAYES	0.6889	0.08679	0.3771	0.7116

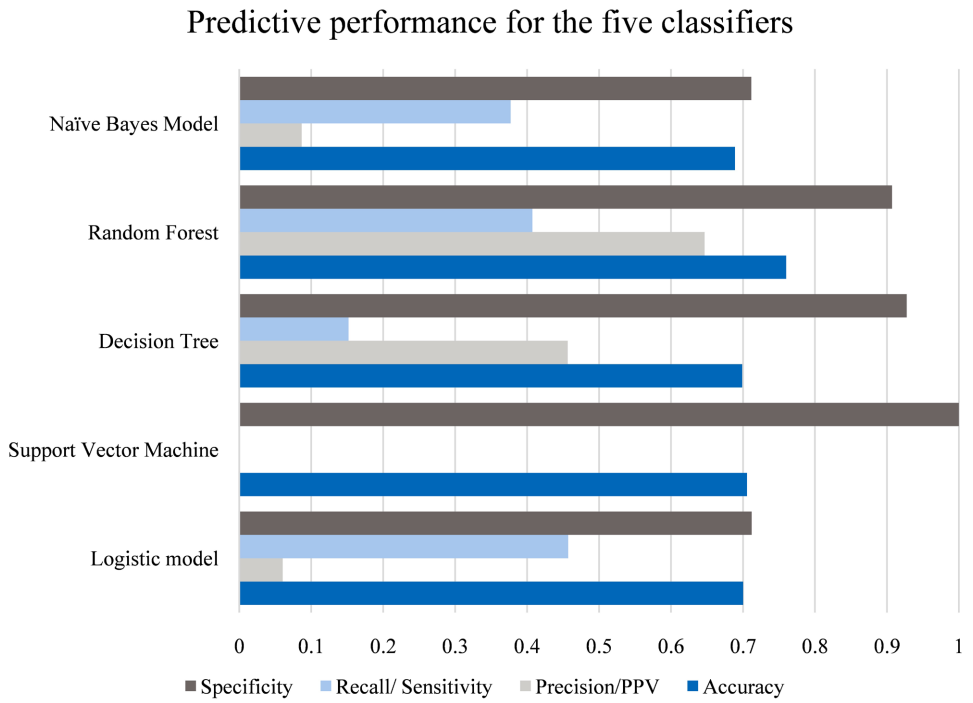

Figure 1. Comparison of the performance metrics across the five classifiers.

The logistic model predicts 616 individuals to have malaria, and this accounts for 6.04% correct prediction. (True Positive = 616, False Positive = 249, precision = 0.06038). Secondly the logistic regression models indicate that the rate of an individual to have malaria to be predicted to fall in the same category by our model is 45.714 percent. (Recall = 0.45714). Specificity of 0.71214, this implies that the model can correctly classify individuals without malaria 71.21% of the time. Accuracy should be as high as possible. It answers the question of how correct our

classifier is generally. The logistic regression model is 70.22% correct (Accuracy = 0.7022) in malaria prediction.

The Support Vector Machine model had both precision and recall of 0, but correctly predicted malaria status of individuals without malaria (specificity = 1.000) with a 70.56% accuracy. The random forest model predicts 576 individuals to have malaria, which accounts for 64.67% correct prediction. (True Positive = 576, False Positive = 157, precision = 0.6467).

The decision tree predicts 589 individuals to have malaria, which accounts for 46.51% correct predictions. (True Positive = 589, False Positive = 225, precision = 0.46512). Secondly the decision tree classifier indicates that the likelihood of an individual suffering from malaria to be predicted as having malaria is 15.09% (Recall = 0.15094). Specificity of 0.9276 implies that our model can predict those without malaria correctly with 92.76% of the time. Accuracy should be as high as possible as it shows how correct the classifier is generally. The decision tree is 69.89% correct (Accuracy = 0.6989).

Further, the decision tree classifier indicates that the likelihood that an individual who has malaria is correctly predicted to have malaria by the model is 40.75% (recall = 0.4075). A specificity value of 0.9071 implies that the model correctly predicts absence of malaria in individuals who do not have the disease 90.71% of the time. Accuracy should be as high as possible. It answers the question of how correct our classifier is generally. The decision tree is 76% correct (Accuracy = 0.76).

The Naïve Bayes classifier predicts 597 individuals to have malaria, and this accounts for 8.679 percent correct prediction. (True Positive = 597, False Positive = 38, precision = 0.08679). Secondly, the decision tree classifier indicates that the rate of an individual having malaria is predicted to fall in the same category by our model at 37.71% (Recall = 0.37705). Specificity of 0.71156 implies that our model can predict those without malaria correctly, with a 71.16% prediction rate.

3.3. Goodness of Fit Measures and Feature Importance

The five classifiers were evaluated with respect to how well they fit the malaria sample data based on the kappa coefficient and model accuracy. The higher the Kappa coefficient the more accurate the classification model. A kappa coefficient of 1 is desired, such that the predicted value equals the observed value. The model with the highest accuracy level and Kappa coefficient value is perceived to be the best fitted model. The goodness of fit measures are summarized in **Table 3** and comparative performance of models displayed in **Figure 2**. The Random Forest has the highest accuracy (0.76) with a 95% confidence interval (0.7307, 0.7876) and Kappa coefficient (0.3526) value and it is therefore better than Logistic model, Decision tree, Support vector model and naïve Bayes model. Though the Random Forest had the highest Kappa coefficient, its value is relatively low, indicating only a moderate compatibility between our classifier and the true class labels while controlling for accuracy.

Table 3. Goodness of fit measures.

MODEL	ACCURACY	KAPPA
LOGISTIC MODEL	0.7022	0.0408
RANDOM FOREST	0.7600	0.3526
SUPPORT VECTOR MODEL	0.7056	0.0000
NAÏVE BAYES MODEL	0.6889	0.0347
DECISION TREE	0.6989	0.0977

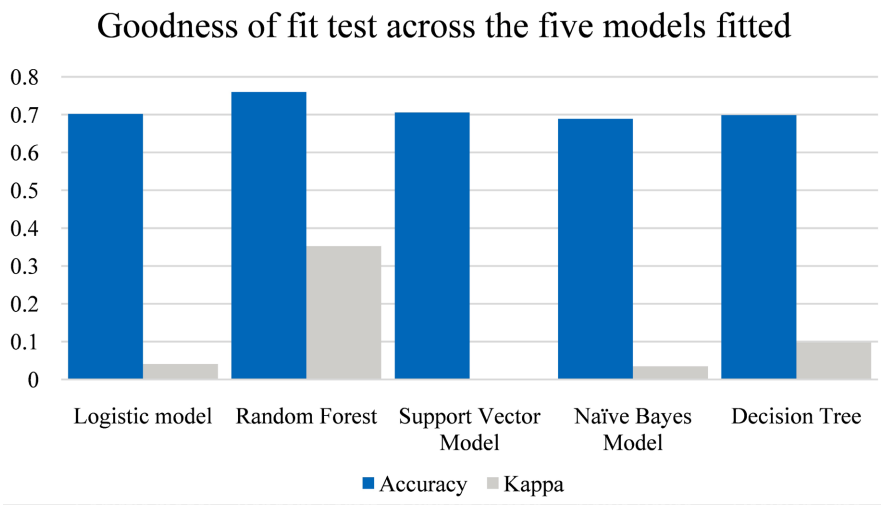

Figure 2. Comparing the goodness of fit across the five classifiers.

McNemar's test was used to compare predictive accuracy of two competing models. A null hypothesis of no difference in predictive ability was tested. The Logistic model, Naïve Bayes model, Support Vector model and Decision tree all had the same McNemar's p-value of 2e−16, while Random forest had a McNemar's p-value of 4.111e−16. Random forest had the highest accuracy and Kappa coefficient value (0.76, 0.3526 respectively). Support vector model came a close second to the Random Forest in terms of accuracy at 71%. Feature Importance and contribution analyses was performed to understand which predictor variables most influenced the Random Forest model. We extracted both Gini importance (computed as the mean decrease in impurity) for each feature, and permutation importance (computed as the average drop in test-fold accuracy when a feature is randomly permuted). The top three predictors by Gini importance were: use of insecticide-treated nets, household wealth quintile and normalized altitude (height above sea level). Permutation importance rankings mirrored this ordering, confirming climate variables as primary drivers with socioeconomic factors also contributing substantially to incidence of malaria.

4. Conclusion and Recommendations

This study has employed five supervised machine learning models, including Logistic regression model, Naïve Bayes model, Support Vector model, Decision tree

and Random Forest, to classify the Kenya Malaria Indicator Survey 2015 data. All five fitted models had relatively high accuracy levels of around 70% with the random forest classifier performing best with 76% accuracy. Support vector model came second at 70.56% followed closely by the logistic model at 70.22%. The precision levels differed from one model to the next with random forest having the highest precision at 64.67%. Decision tree was second highest at 46.51%, while support vector had the least precision at 0%. The analysis also revealed that random forest best fitted the data with an accuracy of 76% and Kappa coefficient value of 0.3526. The foregoing results indicate that Random Forest is the best classifier to predict malaria incidence in individual subjects based on consistently high-performance metrics and goodness-of-fit measures.

Regarding future research, there is need to improve accuracy of the random forest model to above 76%. This could be achieved by streamlining data collection procedures to ensure accuracy and completeness. Improving questionnaire complexity would increase the depth of data collection and provide data on more variables that affect malaria incidence. The kappa coefficient of 0.3526 reflects a fair agreement between climatic and non-climatic variables. Further studies should be carried out to establish the relationship between these climatic and non-climatic factors (with inclusion of additional variables like rainfall, temperature and humidity) to predict malaria. A malaria-free Kenya is possible if accurate forecasting of the case rate is done to enable decision makers to intervene months before onset of any outbreak and potentially save lives.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (WHO) (2025) Malaria. https://www.who.int/news-room/fact-sheets/detail/malaria

[2] Kenya National Bureau of Statistics (KNBS) (2016) Kenya Malaria Indicator Survey 2015. https://www.knbs.or.ke/kenya-malaria-indicator-survey-2015/

[3] Elnour, Z., Grethe, H., Siddig, K. and Munga, S. (2023) Malaria Control and Elimination in Kenya: Economy-Wide Benefits and Regional Disparities. *Malaria Journal*, **22**, Article No. 117. https://doi.org/10.1186/s12936-023-04505-6

[4] Ministry of Health State Department for Public Health and Professional Standards (2025, March 8) Achievement of Indoor Residue Spraying (IRS) in Busia and Seasonal Malaria Chemoprevention (SMC) in Turkana [Press Release].

[5] Guissou, E., Waite, J.L., Jones, M., Bell, A.S., Suh, E., Yameogo, K.B., *et al.* (2021) A Non-Destructive Sugar-Feeding Assay for Parasite Detection and Estimating the Extrinsic Incubation Period of *Plasmodium falciparum* in Individual Mosquito Vectors. *Scientific Reports*, **11**, Article No. 9344. https://doi.org/10.1038/s41598-021-88659-w

[6] Shaw, W.R., Holmdahl, I.E., Itoe, M.A., Werling, K., Marquette, M., Paton, D.G., *et al.* (2020) Multiple Blood Feeding in Mosquitoes Shortens the *Plasmodium falciparum* Incubation Period and Increases Malaria Transmission Potential. *PLOS Pathogens*, **16**, e1009131. https://doi.org/10.1371/journal.ppat.1009131

[7] Dabaro, D., Birhanu, Z., Adissu, W., Yilma, D. and Yewhalaw, D. (2023) Prevalence and Predictors of Asymptomatic Malaria Infection in Boricha District, Sidama Region, Ethiopia: Implications for Elimination Strategies. *Malaria Journal*, **22**, Article No. 284. https://doi.org/10.1186/s12936-023-04722-z

[8] Saini, M., Ngwa, C.J., Marothia, M., Verma, P., Ahmad, S., Kumari, J., *et al.* (2023) Characterization of *Plasmodium falciparum* Prohibitins as Novel Targets to Block Infection in Humans by Impairing the Growth and Transmission of the Parasite. *Biochemical Pharmacology*, **212**, Article ID: 115567. https://doi.org/10.1016/j.bcp.2023.115567

[9] Eva Amoah, L., Kojo Acquah, F. and Kumi Asare, K. (2024) Malaria—Transmission, Diagnosis and Treatment. IntechOpen.

[10] Hagan, R.W., Didion, E.M., Rosselot, A.E., Holmes, C.J., Siler, S.C., Rosendale, A.J., *et al.* (2018) Dehydration Prompts Increased Activity and Blood Feeding by Mosquitoes. *Scientific Reports*, **8**, Article No. 6804. https://doi.org/10.1038/s41598-018-24893-z

[11] Neafsey, D.E., Taylor, A.R. and MacInnis, B.L. (2021) Advances and Opportunities in Malaria Population Genomics. *Nature Reviews Genetics*, **22**, 502-517. https://doi.org/10.1038/s41576-021-00349-5

[12] Erlank, E., Koekemoer, L.L. and Coetzee, M. (2018) The Importance of Morphological Identification of African Anopheline Mosquitoes (Diptera: Culicidae) for Malaria Control Programmes. *Malaria Journal*, **17**, Article No. 43. https://doi.org/10.1186/s12936-018-2189-5

[13] Karim, A.M., Yasir, M., Ali, T., Malik, S.K., Ullah, I., Qureshi, N.A., *et al.* (2021) Prevalence of Clinical Malaria and Household Characteristics of Patients in Tribal Districts of Pakistan. *PLOS Neglected Tropical Diseases*, **15**, e0009371. https://doi.org/10.1371/journal.pntd.0009371

[14] Fornace, K.M., Diaz, A.V., Lines, J. and Drakeley, C.J. (2021) Achieving Global Malaria Eradication in Changing Landscapes. *Malaria Journal*, **20**, Article No. 69. https://doi.org/10.1186/s12936-021-03599-0

[15] Neff, E., Evans, C.C., Jimenez Castro, P.D., Kaplan, R.M. and Dharmarajan, G. (2020) Drug Resistance in Filarial Parasites Does Not Affect Mosquito Vectorial Capacity. *Pathogens*, **10**, Article No. 2. https://doi.org/10.3390/pathogens10010002

[16] Abossie, A., Yohanes, T., Nedu, A., Tafesse, W. and Damitie, M. (2020) Prevalence of Malaria and Associated Risk Factors among Febrile Children under Five Years: A Cross-Sectional Study in Arba Minch Zuria District, South Ethiopia. *Infection and Drug Resistance*, **13**, 363-372. https://doi.org/10.2147/idr.s223873

[17] Ochomo, E.O., Milanoi, S., Abong'o, B., Onyango, B., Muchoki, M., Omoke, D., *et al.* (2023) Detection of *Anopheles stephensi* Mosquitoes by Molecular Surveillance, Kenya. *Emerging Infectious Diseases*, **29**, 2498-2508. https://doi.org/10.3201/eid2912.230637

[18] Sen, P.C., Hajra, M. and Ghosh, M. (2019) Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal, J.K. and Bhattacharya, D., Eds., *Emerging Technology in Modelling and Graphics*, Springer, 99-111. https://doi.org/10.1007/978-981-13-7403-6_11

[19] Wang, C., Chen, X., Du, L., Zhan, Q., Yang, T. and Fang, Z. (2020) Comparison of Machine Learning Algorithms for the Identification of Acute Exacerbations in Chronic Obstructive Pulmonary Disease. *Computer Methods and Programs in Biomedicine*, **188**, Article ID: 105267. https://doi.org/10.1016/j.cmpb.2019.105267

[20] Amin, F. and Mahmoud, M. (2022) Confusion Matrix in Binary Classification Prob-

lems: A Step-by-Step Tutorial. *Journal of Engineering Research*, **6**, Article No. 1.

[21] Sathyanarayanan, S. and Tantri, B.R. (2024) Confusion Matrix-Based Performance Evaluation Metrics. *African Journal of Biomedical Research*, **27**, 4023-4031. https://doi.org/10.53555/ajbr.v27i4s.4345