



## Research article

**Malaria prediction model using advanced ensemble machine learning techniques**

Yusuf Aliyu Adamu\*, Jaspreet Singh

GD Goenka University, Sohna, Gurugram, Haryana, India

**ABSTRACT**

Malaria is a life-threatening disease that leads to death globally, its early prediction is necessary for preventing the rapid transmission. In this work, an enhanced ensemble learning approach for predicting malaria outbreaks is suggested. Using a mean-based splitting strategy, the dataset is randomly partitioned into smaller groups. The splits are then modelled using a classification and regression tree, and an accuracy-based weighted aging classifier ensemble is used to construct a homogenous ensemble from the several Classification and Regression Tree models. This approach ensures higher performance is achieved. Seven different Algorithms were tested and one ensemble method is used which combines all the seven classifiers together and finally, the accuracy, precision, and sensitivity achieved for the proposed method is 93%, 92%, and 100% respectively, which outperformed better than machine learning classifiers and ensemble method used in this research. The correlation between the variables used is established and how each factor contributes to the malaria incidence. The result indicates that malaria outbreaks can be predicted successfully using the suggested technique.

**Keywords:** Malaria prediction, Ensemble Machine Learning Algorithms, climatic and non-climatic factors, CART

Received - 16-08-2021, Reviewed - 12/09/2021, Revised/ Accepted- 14/10/2021

**Correspondence:** Yusuf Aliyu Adamu\* ✉ [adamu2021@yahoo.com](mailto:adamu2021@yahoo.com)

GD Goenka University, Sohna, Gurugram, Haryana, India

**INTRODUCTION**

Malaria becomes the deadly disease globally that is transmitted via a bite of an infected mosquito, it is possible to prevent these deaths by identifying people who are at risk and ensuring that they receive proper and sound care whereby World Health Organization classified it under communicable disease [1]. Relationship between climatic factors, malaria incidence and non-climatic factors have been identified and are related to the development and biting ability of the mosquitoes. Factors that contribute more for the malaria incidence are Temperature, Rainfall, and Humidity, Temperature between 20°C to 30°C gives the conducive environment for the vector development below 16°C and above 30°C curtails their ability to survive, humidity above 60% is said to increase the incidence of malaria and rainfall above 50mm monthly brings favourable conditions for the mosquitoes survival while a combination of high rainfall and high temperatures decreases the likelihood of an occurrence.

Malaria has more effect on African Countries due to their limited health facilities, lack of proper medicine and new modern equipment for diagnosis, poor housing settlements that is why when an outbreak occurs, they cannot control it easily. Its prevention is

very challenges to those countries due to economic instability and hardship [2]. Advancement in technology brings about mechanisms used for early detection and prediction to help control diseases spread and make sure the agencies concern has more time to prepare in advance which can save many lives.

Proper prediction is essential to make sure the authority concern takes all the necessary actions to rescue people in vulnerable areas. When good policies are re-implemented, it reduced or eradicates the malaria incidence. Atmospheric factors, non-climatic factors are important in determining the occurrence of malaria outbreaks in the society and many techniques have been used for the prediction purpose that shows some Algorithms outperform better than others depending on the dataset used [3]. Accuracy is required in knowing how, when, and where the malaria outbreak is liable to happen based on the situations on ground to take measures and save lives from the causative agent of death.

Machine learning Algorithms has been used in many areas, however, researchers are still finding ways to optimize the methods. Ensemble approached is used to enhance the accuracy performances of the classifiers [4]. Different classifiers and techniques are

combined in a predictive model to get better accuracy by reducing variances and minimizing bias<sup>[5]</sup>.

In this study, a new malaria prediction technique is developed that uses CART algorithms to model each part after randomly splitting the datasets into subgroups. To alter the weighted aging ensemble classifier, the accuracy-based weighted aging ensemble classifier (AB-WAE) is utilized, resulting in a more accurate and robust model. For evaluation of the newly proposed model, well-known techniques were tested such as k-nearest Neighbors (KNN), support vector machine (SVM), Naïve Bayes, Decision Tree, Random Forest, Adaboost, Gradient Boost, and Logistic Regression which is used as a stacking ensemble and found out that the newly method proposed has more accuracy for the dataset used. The relationship between the variables used has also been identified and how they contributed to the outbreak.

### Related works

Approaches used for the prediction can be statistical modelling, mathematical and machine learning methods<sup>[6]</sup>. All of these techniques, in combination with computational engineering models, are critical in making decisions and forecasts. Machine learning has been used in health care for diagnoses of many diseases like cancer<sup>[7-8]</sup>, fall detection<sup>[9]</sup>, selecting a better treatment for patients<sup>[10]</sup>, pest prediction and disease detection in Agricultural production<sup>[11]</sup>, detection of mental problems as a result of long time usage of social networking site<sup>[12]</sup>, also in marketing for the prediction of stock exchange<sup>[13]</sup>.

Malaria predictions have been conducted in different countries mostly by considering meteorological variables in the regions, in order to forecast malaria incidence in specific geographic areas over time<sup>[14]</sup>. Different predicting techniques apply to the same dataset by analysing the effectiveness of variables, Researchers use different performances measures to compare methods and models which will provide more prediction accuracy than others for malaria prediction<sup>[15]</sup>.

Machine learning Classifiers that do classification are widely utilised in a variety of field and researchers are continually developing new methods to improve classification performances, one of which is ensemble learning which combines multiple models either heterogeneous or homogeneous using the technique of stacking and Random Forest<sup>[16]</sup>, bagging<sup>[17]</sup> and boosting<sup>[18]</sup>. Numerous medical diagnostic problems have utilized ensemble learning<sup>[19-20]</sup>. Performance can be improved when the ensemble technique is utilized. Many researchers have suggested multiple ways to make ensemble learning by trying to utilized models in different ways to reduce variance and minimize bias. Voting techniques has impacts on the Algorithms performance used in the dataset which has different level of difficulty in analysing the effect of voting

techniques<sup>[21]</sup>. Majority voting is considered usually when using the homogeneous method, some ensemble techniques outperform better than others. According to some research random forest, boosting, and randomized trees performs better than bagging<sup>[22]</sup>. Also, random partitions of the dataset in the ensemble classifier result achieved more improve boosted compatibility than that of the individuals' model<sup>[23]</sup>. An enhanced ensemble model for malaria outbreaks predictions is built in this study.

## MATERIALS AND METHODS

### Dataset Analysis

Malaria dataset used in this research obtained from a database maintained by the World Health Organization ([apps.who.int/gho/data/node.gswcah](https://apps.who.int/gho/data/node.gswcah), [kaggle.com/lydia70/malaria-in-africa](https://kaggle.com/lydia70/malaria-in-africa)) and the weather dataset from world climatic database ([climateknowledgeportal.worldbank.org/download-data](https://climateknowledgeportal.worldbank.org/download-data)).

This dataset was formed to examine malaria outbreak, identify country with highest number of cases and how incidence spread yearly wise, Relationship between variables, identify classifier that perform better during prediction and used for testing the proposed method. The dataset features that contribute the most to the prediction variable were selected. The dataset contains parameters like; Country Name, reported cases of malaria, incidence of malaria in the population's percentage, average Rainfall, Average Temperature, Peoples using basic drinking water services, Latitude, Longitude and basic sanitation services of the population.

Figure 1. Malaria Incidence of the Country

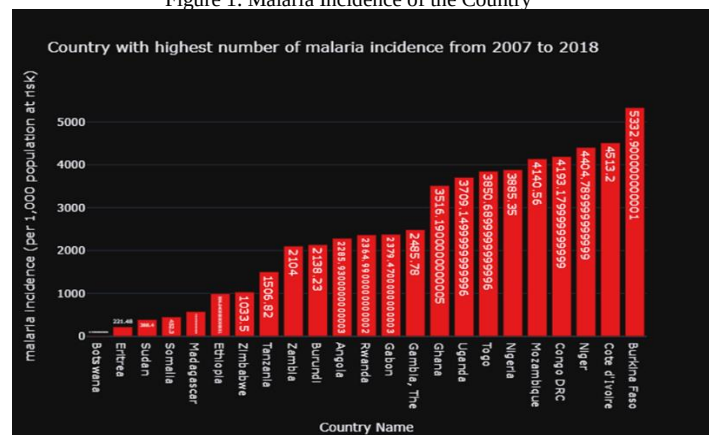


Figure 2. Reported Cases of Malaria Country Wise

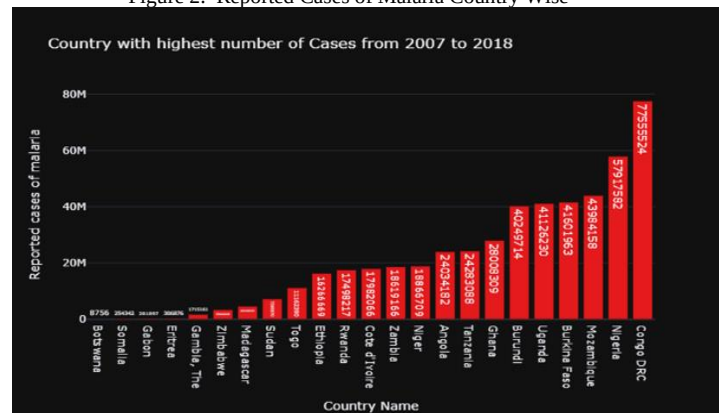


Figure 3: Percentage of Reported Cases Distributed Among the Countries

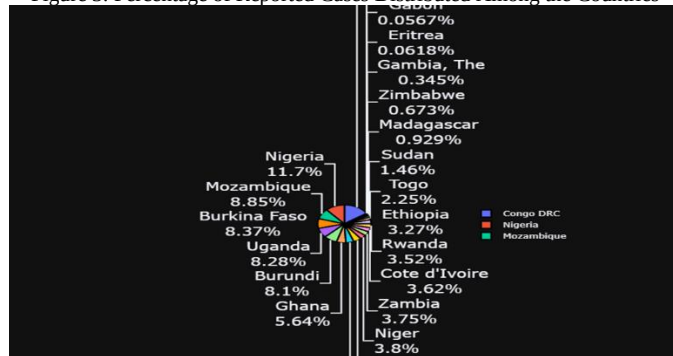


Figure 4: Reported Cases of Malaria with the Country

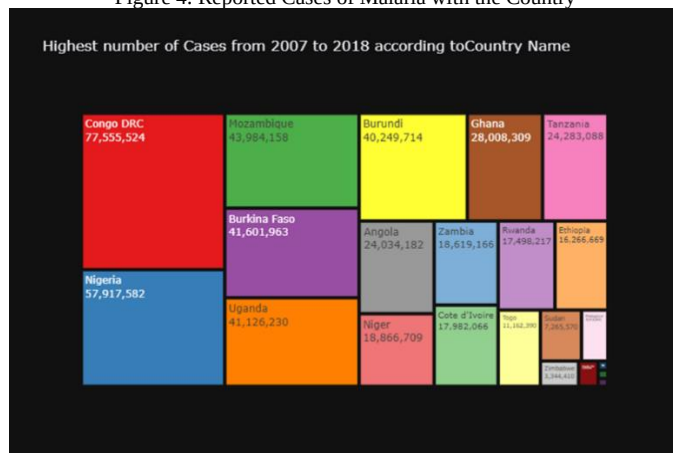
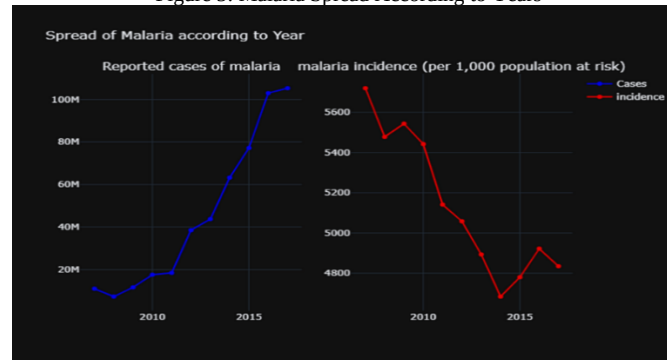


Figure 5: Malaria Spread According to Years



### Parameters Analysis

#### Algorithm used

#### Logistic Regression

This algorithm is used to predict discrete values or categorical variables in which the output takes two values either 0 or 1, yes or no, True or False. It can also be used to forecast discrete values and handle classification tasks. Also used to classify observations using different types of data and easily determines the most effective variables used for the classification. S-shape is used to fit in predicting values of 0 or 1. The logistic regression is derived from the linear regression equation where the y-value ranges between 0 and 1. Written as:

$$\text{Log}\left[\frac{y}{1-y}\right] = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \dots + a_nx_n \quad (1)$$

Figure 6: Comparison of the Parameters through Correlation

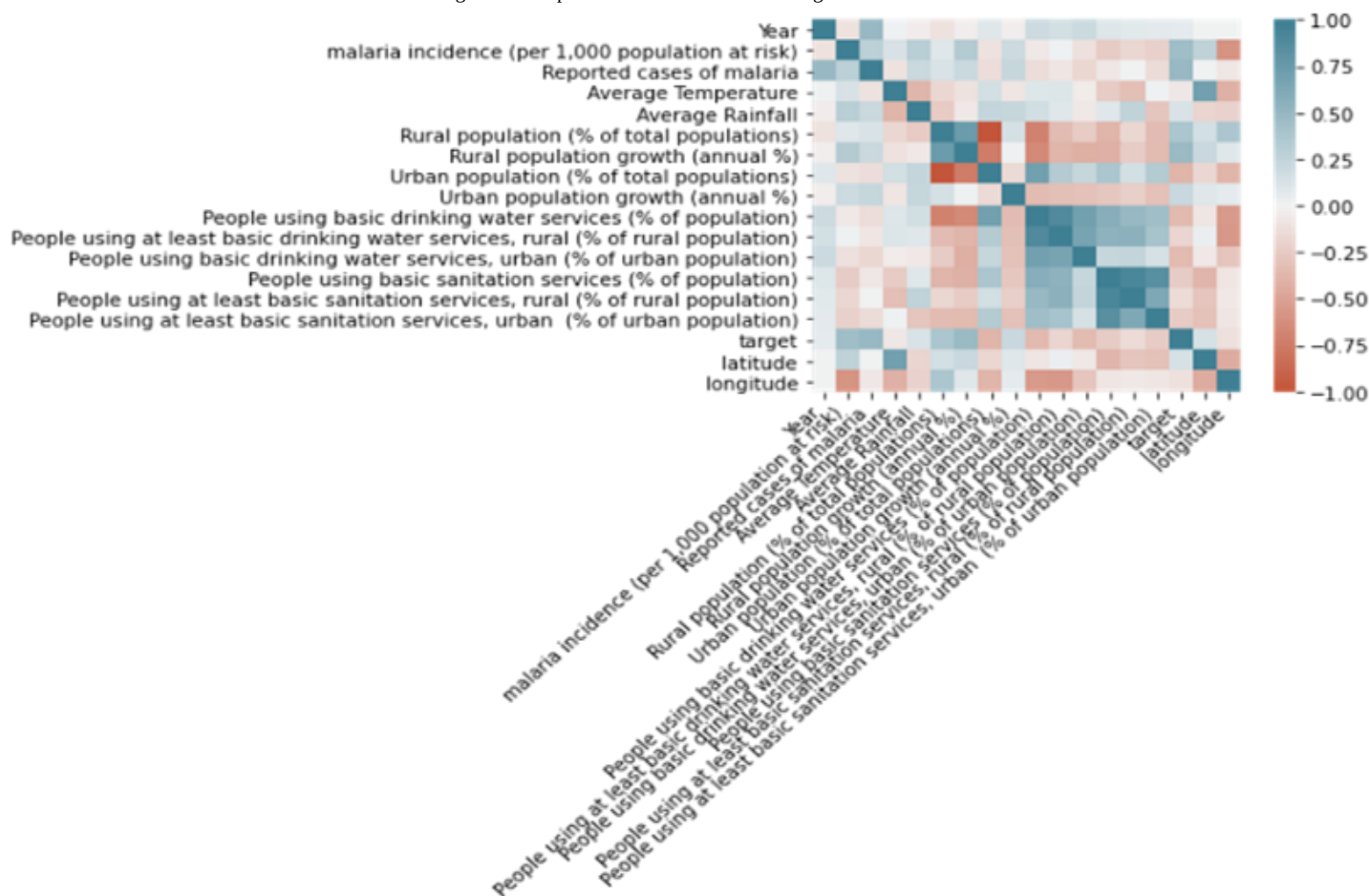
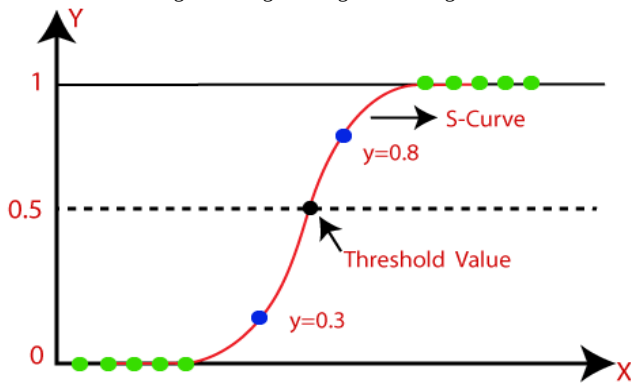


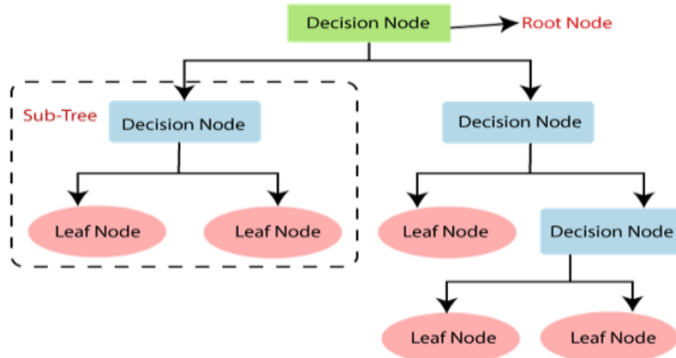
Figure 7: Logistic Regression Diagram



### Decision Tree Algorithm

This algorithm is used to solve both classification and regression problems; it's a tree-like arrangement with nodes and branches that begins from the root node and extends to the leaf node, and it can function with both categorical and continuous variables. The internal node displays the dataset's features, while the branches show the decision rules and the leaf node reflects the problem's result. Cart Algorithm is used in building the tree which it uses decision by answering either yes/no then it further splits the tree into subtrees.

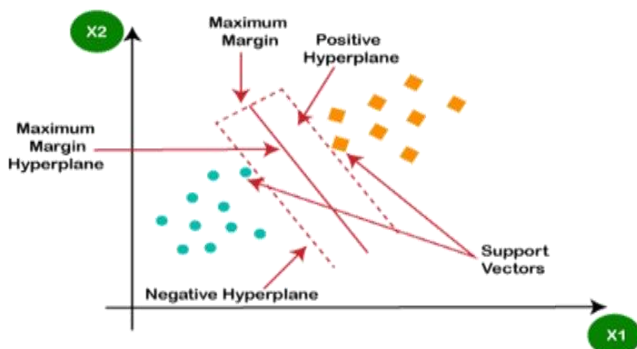
Figure 8: How decision Tree works



### Support Vector Machine

This Classifier applies to both classification and regression matters; Its purpose is to develop a hyperplane data point that can classify datasets into distinct groups. Support vector is a data point that defines the hyperplane. [28].

Figure 9: SVM Diagram



### Naïve Bayes Algorithm

This classifier is used to produce prediction based on the object's probability in such a way that variables are independent to each other.

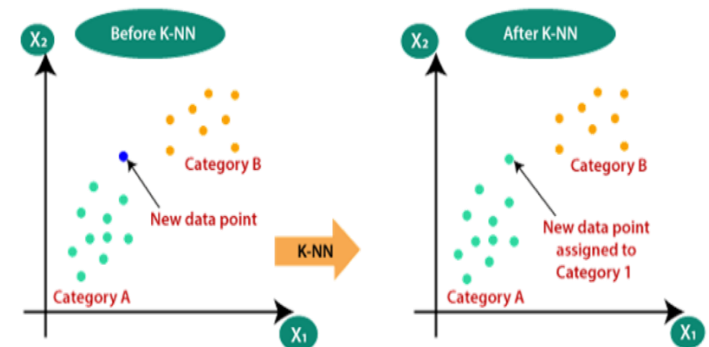
$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad \text{----- (2)}$$

### K-Nearest Neighbors (KNN)

Both classification and regression issues are solved using this method. It works on the basis of assumptions, in which the similarity between new data points and existing data points is checked, and new data points are placed in the most comparable categories based on these similarities. It's also known as the lazy learner algorithm since it keeps all of the data during the training stage and then classifies new data into appropriate classes when it comes in. This means it doesn't learn from the training set right away. It uses Euclidean formula to find the distance between classes. Written as:

$$\text{Distance between two points} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad \text{----- (3)}$$

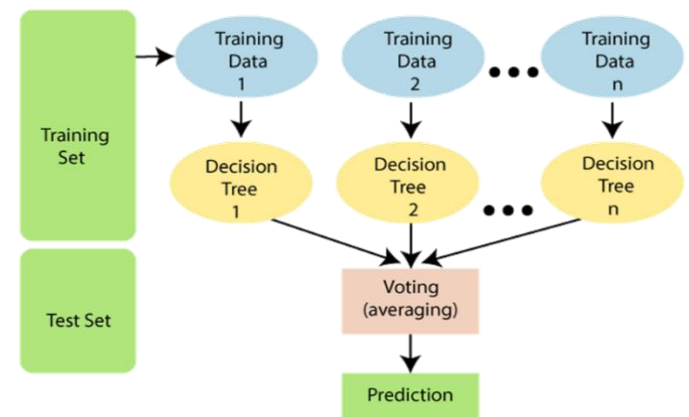
Figure 10: Diagram of How KNN Algorithm Works



### Random Forest

This is one of the rapid, efficient and powerful supervised learning Algorithm that contain multiple decision trees and both classification and regression problems can be solved using this method. It employs an ensemble approach by combining multiple classifiers to improve the model performance during predictions.

Figure 11: How Random Forest Work



### AdaBoost Classifier

This Algorithm used to fit a series of weak base learners on modified version of data repeatedly. To get the given output, it uses a weighted majority vote to integrate the predictions of all the base learners. It involves applying weights to each of the training dataset.

### Gradient Boosting Algorithm

This is an effective and correct method used to tackle classification and regression issues. It involves ensemble of weak models' prediction more especially decision tree which it allows optimization of the arbitrary differentiable loss function and the model is built in a stage-wise like other boosting methods.

### Stacking Classifier

Stacking is a machine learning approach that combines predictions from numerous base learners to create a final prediction that outperforms any single model in the ensemble. Meta model uses the output from the base learners as inputs in the case of regression, the outputs from the base models are real values, while in the case of classification, the outputs are probability values. It uses a K-fold cross-validation approach for dataset training preparation.

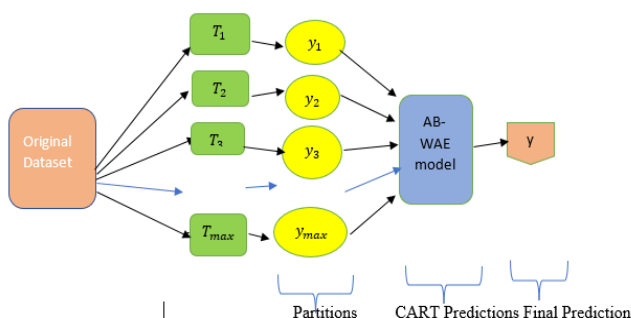
### Proposed Technique

Some of the methods applied in this work include:

(a) Naive Bayes (b) K-Nearest Neighbors (c) Support Vector Machine (d) Random Forest (e) Decision Tree (f) Adaboost classifier (g) Gradient Boost classifier (h) Stacking Technique (i) New proposed model. This research takes a purely data-driven approach whereby seven classification Algorithms are employed to assess the accuracy, precision and sensitivity of the predictions. Stacking ensemble technique is used to improve the accuracy and lastly, a new technique developed for the prediction and its result is compare with all other classifiers.

The diagram below shows how the dataset is partitioned into multiple groups from  $T_1$  to  $T_{max}$ , a predicted values were obtained using CART from  $y_1$  to  $y_{max}$  which allows the computation of the ensemble technique from the various CART models, the accuracy-based weighted ageing classifier ensemble (AB-WAE) is utilised to get the final prediction.

Figure 12: Methodology of the Newly Proposed Method



A new proposed method for making a prediction is shown in figure 12. Let assume the dataset be represented as D,

where  $D = \{(X_i + Y_i), \text{ and } i = 1, 2, 3, 4, \dots, M\}$ .----- (4)

The predictor variables X can be written as  $X_i = [X_{i1}, X_{i2}, X_{i3}, X_{i4} \dots X_{iN}]$  and the response variable Y in the dataset such that  $y_i \in \{0, 1\}$ . Looking at the individual independent variables of different instances or rows represented as:

$$X_j = [X_{1j}, X_{2j}, X_{3j}, X_{4j}, \dots, X_{Mj}] \text{ ----- (5)}$$

The weighted mean of the variable  $X_j$  can be computed using

$$\bar{X}_j = \sum_{i=1}^m Q_i X_{ij}, \bar{Q}_i \geq 0, \sum_{i=1}^m \bar{Q}_i = 1 \text{ ----- (6)}$$

From the above equation, it indicates that the weighted mean is influenced more by data elements having a high weight than that with a low weight. The weights can't be negative, despite that some value can be 0.

Randomization is used in the creation of the proposed data partitioning technique. The algorithm pulls at random in the set of variables,  $\{1, 2, 3, \dots, N\}$  at each partitioning. This randomization can generate any number of trees, reducing variation and improving performance [29].

Consider the root node, which denote the entire class that can be divided into distinct homogenous classes. Choose a point of data at random from the variable,  $\{1, 2, 3, \dots, N\}$  and name it j. This point of data is removed from the set to allow more options. Using the mean-based partitioning rule, divide D into two partitions. Consider each child node separately  $D_{11}$  and  $D_{12}$  as root nodes and by applying equation 7 above. The dataset  $D_{11}$  creates  $D_{21}$  and  $D_{22}$  and  $D_{12}$  creates  $D_{23}$  and  $D_{24}$ .

$$D = \begin{cases} D_{11}, & \text{if } X_{ij} < \bar{X}_j \\ D_{12}, & \text{if } X_{ij} \geq \bar{X}_j \end{cases} \text{ ----- (7)}$$

This process of partitioning will continue until the stopping rules are reached. In this new technique, to avoid over partitioning the dataset, two restrictions are employed. The minimum tree height is the first criterion, which prevents the algorithm from dividing the tree when it reaches a certain height at  $H = TH_{max}$ . The tree height is defined as  $H = 0$  for the root node and  $H = 1$  for  $D_{11}$  and  $D_{12}$  continuously. In the second rule, the tree will remain inactive if the partition has a small number of instances.

Let assume  $M(D_{ki})$  represent the number of instances in node  $D_{ki}$ , this process will stop when  $M(D_{ki}) < M_{min}$ , in which  $M_{min}$  is the partition's intended minimum instance count. As a result, the tree reaches a point where it can no longer grow if,

$$H = TH_{max} \text{ or } M(D_{ki}) < M_{min} \text{ ----- (8)}$$

Following the portioning of the dataset, because decision tree models offer a high level of accuracy and reliability, each partition is modelled using a classification and regression tree. When a specific variable is picked at random, CART uses Gini Impurity to assess the likelihood of it being erroneously classified.

To determine the Gini Impurity for a group of J-class elements. Use the following formula:

Let say  $j \in \{1, 2, \dots, J\}$

$$\text{Gini} = 1 - \sum_{i=1}^J P_i^2 \text{ ----- (9)}$$

Where  $P_i$  denotes the likelihood of an object being

categorised into a specific category,  $T_{max}$  represent the forest of tree which are all fitted with CART Algorithm. To compute the ensemble from the various CART models, the accuracy-based weighted ageing classifier ensemble (AB-WAE) is utilised. The weighted aging classifier ensemble has been modified to create the AB-WAE. [25].

Let say  $P_{\alpha}(\beta_i)$  indicates the proportion accurate responses of classifier  $\beta_i$  while  $itter(\beta_i)$  represent the number of times  $\beta_i$  has changed in the ensemble. Therefore, the classifier weight  $w(\beta_i)$  can be written as:

$$P_{\alpha}(\beta_i) > P_{\alpha}^{\pi} \text{ then } w(\beta_i) = P_{\alpha}(\beta_i) \text{ (Psii)}$$

$$\text{Else, } w(\beta_i) = \frac{P_{\alpha}(\beta_i)}{\sum_{j \in \{1,2,\dots\}} P_{\alpha}(\beta_j)} \text{ -----(10)}$$

Such that  $P_{\alpha}^{\pi}$  reflect the ensemble's classifier's average accuracy  $\pi$ . The ultimate prediction can be obtained as:

$$\beta(x) = i \text{ if } \sum_{t=1}^{T_{max}} w(\beta_i) F_t^i(x) = \max_{j \in \{1,2,\dots\}} \sum_{t=1}^{T_{max}} w(\beta_j) F_t^j(x) \text{ ----- (11)}$$

Classifiers are given weights based on the time taken and the accuracy to complete the ensemble. A classifier is eliminated from the ensemble if its weight falls below a specific threshold [26]. The ensemble gets perfect outcomes when the accuracy is used as the optimization criterion [30].

### Model Evaluation

Python code is used to calculate the prediction of the malaria dataset to determine accuracy, precision, recall, error rate, sensitivity macro, and weighted average of different classification techniques. The accuracy of any classification method is calculated using a confusion matrix and is defined as:

**True Positive (TP)** - Positive target variables that are predicted as Positive correctly.

**False Positives (FP)** –Total numbers of negative variables that are Predicted as positive wrongly.

**True Negatives (TN)** – Negative Variables targets that are predicted Correctly as negative.

**False Negatives (FN)** - Positive variables targets that are predicted as negative wrongly

Accuracy is the rate of correct classifications

$$\text{Accuracy} = \frac{\text{Number of correct predictions made}}{\text{Total number of predictions made}} \text{ ----- (12)}$$

Precision: Is the proportion of true positive outcome to positive outcome anticipated.

$$\text{Precision} = \frac{TP}{TP+FP} \text{ -----(13)}$$

Sensitivity: The proportion of correct positive forecasts to total positive cases.

$$\text{Recall} = \frac{TP}{TP+FN} \text{ -----(14)}$$

F1-score: Is the harmonic mean of accuracy and sensitivity

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \text{ -----(15)}$$

### RESULT

Three types of experiments are carried out in this section to assess the efficacy of the proposed method. To begin, the performance of individual classifiers is determined by employing seven well-known Machine Learning Algorithms; support Vector Machine, naïve Bayes, K-Nearest Neighbours, Decision Tree, Random Forest, Ada boost classifier, and Gradient Boost classifier. Secondly, a Logistic Regression is used as a stacking classifier to enhanced the accuracy performance of the Algorithms which combined all seven classifiers. Lastly, the method proposed was tested using malaria datasets and compare its classification accuracy to those of other machine learning methods

Table 1: Performance of Individual Classifiers

| Algorithms                | Accuracy (%) | AUC (%) | Precision (%) | Recall (%) | F1 score (%) | Std Mean(+/-) | Specificity | Error Rate(%) | Macro Average | Weighted Average |
|---------------------------|--------------|---------|---------------|------------|--------------|---------------|-------------|---------------|---------------|------------------|
| SVM                       | 85.2         | 85      | 77            | 85         | 81           | 0.046         | 93          | 12.5          | 85            | 88               |
| KNN                       | 85.7         | 87      | 73            | 80         | 76           | 0.034         | 90          | 15.63         | 82            | 85               |
| Random Forest             | 91.6         | 90      | 83            | 100        | 91           | 0.042         | 100         | 6.25          | 92            | 95               |
| Decision Tree             | 90.5         | 90      | 86            | 95         | 90           | 0.072         | 98          | 6.25          | 92            | 95               |
| Naïve Bayes               | 84.1         | 88      | 71            | 100        | 83           | 0.056         | 100         | 12.5          | 86            | 91               |
| AdaBoost Classifier       | 85.2         | 80      | 87            | 100        | 93           | 0.046         | 100         | 4.69          | 93            | 96               |
| Gradient Boost Classifier | 88.4         | 88      | 87            | 100        | 93           | 0.033         | 100         | 4.69          | 93            | 96               |

The significance performance of one classifier over another is shown in table 1, which indicates, Random Forest classifier has the

highest performance in accuracy than all other classifiers that are involved in the analysis.

Table 2: Ensemble Method and New Proposed Technique

| Algorithms          | Accuracy (%) | AUC (%) | Precision (%) | Recall (%) | F1 score (%) | Std Mean(+/-) | Specificity | Error Rate(%) | Macro Average | Weighted Average |
|---------------------|--------------|---------|---------------|------------|--------------|---------------|-------------|---------------|---------------|------------------|
| Stacking Classifier | 92.1         | 91      | 83            | 96         | 89           | 0.059         | 97          | 9.37          | 90            | 92               |
| Proposed Method     | 93           | 90      | 92            | 100        | 93           | 0.051         | 100         | 4.55          | 93            | 96               |

Table 2, as shown it represent the Ensemble of all the seven base learners' classifiers performances using Logistic Regression by applying stacking technique of ensemble method, to enhanced the accuracy and lastly the effectiveness of the newly presented method. To confirm the effectiveness and accuracy of the proposed technique,

Analysis is done on all the classifiers in table 1 and 2, outlined the test results of multiple classifiers on the datasets used, it is observed that the stacking ensemble learning method performed better than individual classifier and the proposed method out performed all other classifiers in terms of performance

## CONCLUSION

Machine learning techniques are crucial in the prediction of malaria outbreaks and can be used to create a model that can help forecast various forms of illness. Malaria is a greatest cause of mortality more particular in developing countries. Timely identification can aid the prevention of disease development.

There is clear evidence that rainfall, Temperature, Environmental sanitation and basic drinking water plays an important role in malaria incidence; thus, proper sanitation is required to avoid malaria incidence in a community.

Various data mining techniques for malaria outbreak prediction were identified. The following machine-learning techniques are used: Nave Bayes, Support Vector Machine, Decision Tree, Random Forest, k-Nearest Neighbors, Adaboost and Gradient Boosting. Logistic Regression was employed as a stacking classifier to enhance the accuracy of the individual classifiers, and it was discovered that the new approaches give more effectiveness than that of the ensemble method.

The methodology allows spontaneously splitting the data into smaller units using a mean-based splitting algorithm, followed by classification and regression to model every division to correctly predict malaria outbreaks. The suggested method obtained accuracy, precision, and sensitivity of 93%, 92%, and 100%, respectively. In comparison to previous machine learning approaches suggested, the proposed ensemble obtained showed improved performance.

## REFERENCES

- W Cella, 2019. Do climate changes alter the distribution and transmission of malaria? Evidence assessment and recommendations for future studies Rev. Soc. Bras. Med. Trop.
- A. Midekisa, B. Beyene, A. Mihretie, E. Bayabil, M. C. Wimberly, 2015. Seasonal associations of climatic drivers and malaria in the highlands of Ethiopia.
- World Health Organization, 2020. Malaria Report.
- Yusuf Aliyu, Adamu, Singh J, 2021. Malaria Prediction Model Using Machine Learning Algorithms. Turkish Journal of Computer and Mathematics Education, Vol.12 No.10, 7488-7496. Doi-10.17762/turcomat.v12i10.5655.
- Sevakula RK, Verma NK, 2017. Assessing generalization ability of majority vote point classifiers. IEEE Transactions on Neural Networks and Learning Systems. 28(12), 2985-97. Doi-10.1109/TNNLS.2016.2609466.
- Li H, 2018. Ensemble learning for overall power conversion efficiency of the all organic dye-sensitized solar cells. IEEE Access 6:34118-26. Doi- 10.1109/ACCESS.2018.2850048.
- K., Kigozi, R., Charland, K., Dorsey, G., Kamya, M., Buckeridge, D, 2013. Predicting Malaria in a Highly Endemic Country using Environmental and Clinical Data Sources. Online journal of public health informatics, 6(1).
- Danger, R., Segura-Bedmar, I., Martínez, P., Rosso, P, 2010. A comparison of machine learning techniques for detection of drug target articles. Journal of biomedical informatics, 43(6), 902-913.
- Urquiza, J. M., Rojas, I., Pomares, H., Herrera, J., Florido, J. P., Valenzuela, O., Cepero, M, 2012. Using machine learning techniques and genomic/proteomic information from known databases for defining relevant features for PPI classification. Computers in biology and medicine, 42(6), 639-650.
- Rastogi S, Singh J, 2012. A systematic review on machine learning for fall detection system. Computational Intelligence.37:951- 974. Doi- 10.1111/coin.12441.
- Worner, S. P., Gevrey, M, 2006. Modeling global insect pest species assemblages to determine risk of invasion. Journal of Applied Ecology, 43(5), 858-867.
- Shashi, Singh, J, 2021. Automated Disease Detection and Classification of Plants Using Image Processing Approaches. A Review. In P. K. Singh, S. T. Wierzchoń, S. Tanwar, M. Ganzha, and J. J. P. C. Rodrigues (Eds.). Proceedings of Second International Conference on Computing, Communications, and Cyber-Security, pp. 641-651. Springer. Doi-10.1007/978-981-16-0733-2\_45
- Singh, A., Singh, J, 2021. Automation of detection of social network mental disorders. A review. IOP Conference Series: Materials Science and Engineering, 1022, 012008. Doi-10.1088/1757-899X/1022/1/012008
- Patel, J., Shah, S., Thakkar, P., Kotecha, K, 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259- 268.
- Zinszer, K., Kigozi, R., Charland, K., Dorsey, G., Kamya, M., Buckeridge, D, 2013. Predicting Malaria in a Highly Endemic Country using Environmental and Clinical Data Sources. Online journal of public health informatics, 6(1).
- Abisoye, Opeyemi A., Jimoh Gbenga R, 2017. Symptomatic and Climatic Based Malaria Threat Detection Using Multilevel Thresholding Feed Forward Neural Network. I.J. Information Technology and Computer Science, 8, 40-46
- Ho Tin Kam, 1998. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 20(8):832-44. <https://doi.org/10.1109/34.709601>.
- Breiman L, 1996. Bagging predictors. Mach Learn 24(2):123-40. Doi-10.1023/A:1018054314350.
- Schapire RE, Singer Y, 1999. Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297-336. Doi- 10.1023/A: 1007614523901.
- Zhang B, 2019. Ensemble learners of multiple deep CNNs for pulmonary nodules classification using CT images. IEEE Access 7:110358-71. Doi- 10.1109/ACCESS.2019.2933670.
- Han L, Luo S, Yu J, Pan L, Chen S, 2015. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE Journal of Biomedical and Health Informatics. 19(2):728-34. Doi-10.1109/JBHI.2014.2325615.
- Leon F, Floria S-A, Badica C, 2017. Evaluating the effect of voting methods on ensemble based classification. IEEE international conference on Innovations in intelligent Systems and applications (INISTA). p. 1-6. Doi-10.1109/INISTA.2017.8001122.
- Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP, 2007. A comparison of decision tree ensemble creation techniques. IEEE Trans Pattern Anal Mach Intell 29 (1):173-80. Doi-10.1109/TPAMI.2007.250609.

24. Ruta D, Gabrys B, Lemke C, 2011. A generic multilevel architecture for time series prediction. *IEEE Trans Knowl Data Eng.* 23(3):350–9. Doi- 10.1109/TKDE.2010.137.
25. S. Patel, 2018. Machine Learning. Supervised Learning and Naive Bayes Classification online 101,
26. T. Srivastava, 2018. Introduction to k-Nearest Neighbors: Simplified with implementation in Python. *Analytics Vidhya*.
27. N. Donges, 2018. The Random Forest Algorithm. Towards Data Science. Ensemble methods — scikit learn 0.24.2 documentation.
28. Rivera-Lopez R, Canul-Reich J, 2018. Construction of near-optimal axis-parallel decision trees using a differential-evolution-based approach. *IEEE Access* 6:5548–63. Doi-10.1109/ACCESS.2017.2788700.
29. Woźniak M, 2013. Application of combined classifiers to data stream classification. In: Saeed K, Chaki R, Cortesi A, Wierzchon S (editors). *Computer information Systems and industrial management*, vol. 8104. Berlin; Heidelberg, p. 13–23.
30. Wozniak M, 2017. Accuracy based weighted aging ensemble (AB-WAE). Algorithm for data stream classification. In: *IEEE 4th international Conference on soft computing machine intelligence (ISCMI)*, p. 21–4. doi 10.1109/ISCMI.2017.8279591.

#### How to cite this article

Yusuf Aliyu Adamu, Jaspreet Singh, 2021. Malaria prediction model using advanced ensemble machine learning techniques. *Jour. of Med. P<sup>r</sup>actical & Allied. Sci.* V 10 - I 6, 1701, P- 3794-3801. doi: 10.22270/jmpas.V10I6.1701