MALARIA PREDICTION MODEL USING MACHINE LEARNING ALGORITHMS

Yusuf Aliyu Adamu[a] and Jaspreet Singh[b].

[a]Ph.D. Research Scholar Faculty of Engineering, GD Goenka University, Sohna, India
[b]Faculty of Engineering, GD Goenka University, Sohna, India

ABSTRACT
Measures have been taking to ensure the safety of individuals from the burden of vector-borne disease but it remains the causative agent of death than any other diseases in Africa. Many human lives are lost particularly of children below five years regardless of the efforts made. The effect of malaria is much more challenging mostly in developing countries. In 2019, 51% of malaria fatality happen in Africa which it increased by 20% in 2020 due to the covid-19 pandemic. The majority of African countries lack a proper or a sound health care system, proper environmental settlement, economic hardship, limited funding in the health sector, and absence of good policies to ensure the safety of individuals. Information has to become available to the peoples on the effect of malaria by making public awareness program to make sure people become acquainted with the disease so that certain measure can be maintained. The prediction model can help the policymakers to know more about the expected time of the malaria occurrence based on the existing features so that people will get to know the information regarding the disease on time, health equipment and medication to be made available by government through it policy. In this research weather condition, non-climatic features, and malaria cases are considered in designing the model for prediction purposes and also the performance of six different machine learning classifiers for instance Support Vector Machine, K-Nearest Neighbour, Random Forest, Decision Tree, Logistic Regression, and Naïve Bayes is identified and found that Random Forest is the best with accuracy (97.72%), AUC (98%) AUC, and (100%) precision based on the data set used in the analysis.
Keywords: Malaria prediction, Machine Learning Algorithms, climatic and non-climatic factors.

1.0 INTRODUCTION

Malaria is a dangerous communicable disease that leads to the death of peoples in the world every day despite the effort makes for its eradication since it can be prevented if certain measures are putting in place but remain the causative agent of death more especially in Africa because of its impact in human lives [1]. In 2019 worldwide malaria statistics show that 94% of the cases and 51% of death happened in Africa whereby Nigeria has the highest rates followed by the Republic of Congo (DRC), Republic of Tanzania, Republic of Mozambique, Niger, and Republic of Burkina Faso regarded as 23%, 11%, 5%, 4%, 4%, and 4% respectively with resultant death of 409,000 globally [2].

Vector-borne disease effects are more in Africa because of a limited medical resource, lack of proper medication, inadequate equipment, inefficient funding, lack of policies to manage the situation, economic hardships, poor environmental sanitation, and housing conditions [2]. In the year 2020, Malaria cases in Africa increase because of coronavirus pandemic becomes a challenge that leads to the loss of many lives in the world. This is a serious challenge to health care management and non-governmental organizations in finding all the possible measures to bring the long-lasting question on why Malaria becomes a causative agent of death more especially in third world countries.

Malaria is associated as the most weather pattern quick to respond disease [3]; therefore, changes in the atmospheric conditions can help widespread its presence by a direct route or indirectly through human actions via improper cleaning by dumping waste product and industrial waste [4]. Numerous factors such as crowded environmental pattern, existing of unclean gutters, presence of garbage disposers and shallow pools of brackish water which help mosquitos' larvae to complete their growth development.

All report suggests that vulnerable places could be rescued if a good Policy can be put in place to tackle malaria scourge; the number of peoples affected will reduce or eradicated at all. This can be achieved when the policymakers become aware of where the malaria instances are liable to occur all the necessary action can be done to save lives. Therefore, there is a need for a proper prediction mechanism on when, where, and how the outbreak will occur.

This research work used different Algorithms in designing the model that can help to predict the instance of malaria and also help the government, health care providers, and relevant pertinent to take all the possible measures in ensuring the safety of individuals long before the epidemics can be achieved by enlighten peoples on the effect of vector-borne disease and increase malaria prevention facilities [5-7].

2.0 LITERATURE REVIIEW

Multiple researchers make an effort to predict outbreaks by the use of weather conditions which is likely the most foretell of malaria transmission, some factors are also attributed to different circumstances [8,10]. It has been found that heavy or less rainfall is not only the factor that influences the incidence of malaria [11]. Rainfall and temperature are the essential components that contributed to malaria transference [12]. Analysis has shown that the rate of fatality increase during the rainy season regardless of the temperature and humidity [13].

Temperature, precipitation and humidity influence the mosquito’s life cycle for their growth and development [14]. The vector larvae survive only when the environmental condition is conducive in a moderate temperature above 160C and die when it is lower or higher [15]. The rate at which mosquitoes bite humans by sucking their blood increases when the environmental conditions are favourable for their survival.

The approach used for prediction ranges from mathematical to statistical modelling and machine learning algorithms [16]. Mathematical and statistical models play an important role in predictions to make decisions. Machine learning is used in health care and helps diagnose many diseases such as cancer [17-18]. It also helps pharmacologists to find the right formula for forming a reliable medicine to incapacitate a disease virus [19,20]. The machine is also used in selecting fruitful treatment [20]. Also can be used in Agricultural production for predicting pest plants [21]. Also stock exchange in the market can be predicted [22].

Selecting the techniques to use for proper predictions depend on the problems to be solved include Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Artificial Neural Network (ANN), Random Forest, Logistic Regression and Gradient Extreme Boost Algorithm [23].

3.0 **MATERIALS AND METHOD USED**

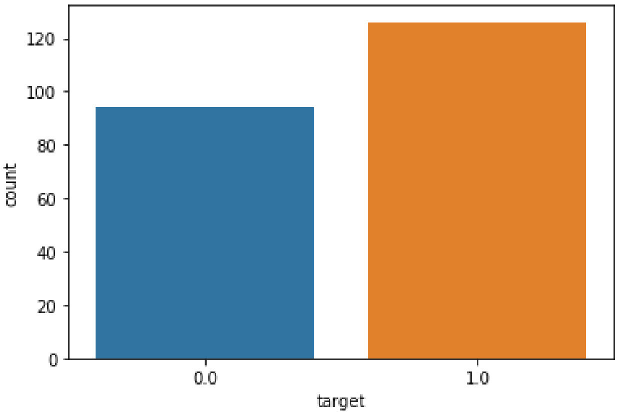
3.1 **Data Collection**

The datasets used to train the machine learning Algorithms to contain different parameters, which includes the percentage of the population using at least basic sanitation, Percentage of the population using at least basic drinking water, Average Temperature, Average Rainfall, Total number of yearly malaria reported cases, Total Incidence of malaria (populations at risk per 1,000).
The malaria incidence dataset was found from the portal of the world health organization http://apps.who.int/gho/data/node.gswcah, https://www.kaggle.com/lydia70/malaria-in-africa and the Atmospheric dataset of Rainfall and Temperature from the world climate portal https://climateknowledgeportal.worldbank.org/download-data.

3.2 Building a Model

Jupyter notebook was used, it can be downloaded free of cost and stored user work in a complete format that contains text and programming codes which is a tool for data analysis and machine learning algorithms are used to analysed data and also used to developed models for making predictions.

Figure 1: Total number of Positive and Negative cases



:

Figure 2: Represent the histogram of the independent variables

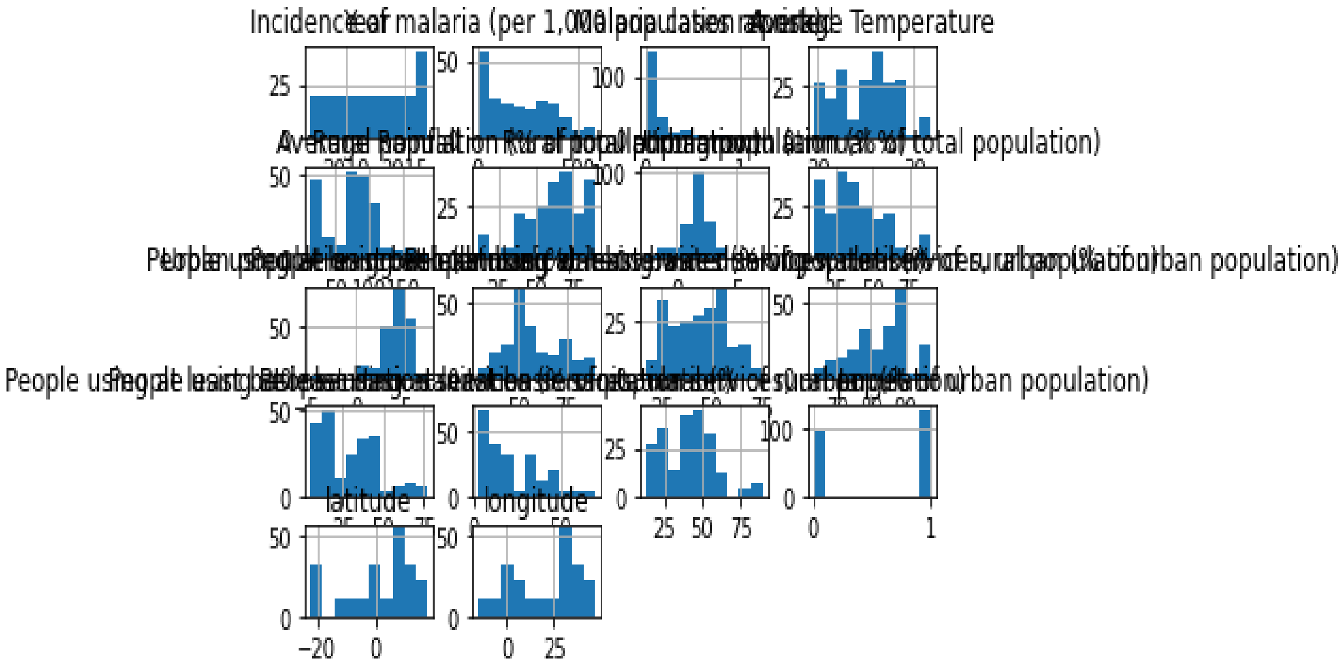
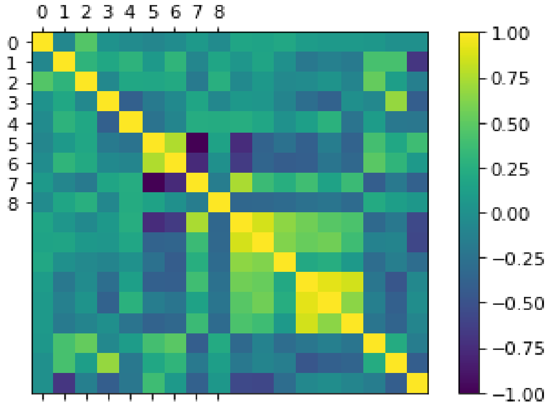


Figure 3: Represent the correlation between variables (Heatmap)



3.3 Data transformation and Execution

Data used here are in CSV file format by integrating the Atmospheric dataset and malaria incidence datasets into a single file structure. The dataset used in this case is split into two categories, eighty Percent (80%) and twenty percent (20%) for the model training and testing respectively. Also, six different classification techniques are used, Support Vector Machine, Logistic Regression, Random Forest Classifier, Naïve Bayes, Decision Tree, and KNN.

3.4 Algorithms Used

Popular classification algorithms were used and compared their performances.

3.4.1 Support Vector Machine

Support Vector Machine (SVM) or Support Vector Classifier (SVC) is used to calculate the hyperplane to separate groups of different classes [24]. In this algorithm, each feature has a corresponding coordinate in the N-Dimensional X, Y plain graph whereby each value represents a particular data point for classification. It is one of the famous techniques.

Let the data to be train be, A

$$\text{Therefore, } A = (BiCi) \mid Bi \in T,\ Ci \in \{-1,1\},\ i = 1 \text{ to } n \qquad (1)$$

The Ci can have to values (-1 and +1), this values determine the position of Bi in the vector class. A hyperplane is a set of points B satisfying, $P \,.\, B – K = 0$, whereby P represent the vector to the hyperplane and $K / \|P\|$ signifies the effect of the vector in finding the maximum distance of Ci.

3.4.2 K-Nearest Neighbors (KNN) Algorithm

It is one of the supervised machine learning techniques that are popular based on its simplicity in solving a classification problem. It uses to predict a new class by identifying its closest neighbors in the observations which makes it have more predictive power and spend less time during interpretation of the output because the majority vote of its k nearest common classes are considered [25]. If the value of K is well selected it will minimize the training and validation time.

3.4.2 Logistic Regression Algorithm

It is when dependant and independent variables are of the same binary category. The probability of each group is estimated since there are no linear relationships between the variables. The final result of the observation belonging to a particular group has to be binary [26].

Z represents a numeric value target variable of linear Regression.

$$P = \frac{1}{1+e^{\mathrm{exp}-Z}} \qquad (2)$$

3.4.4 Random Forest Algorithm

This technique is used to solve almost all the problems either in binary or Regression; it works by combining several decision trees [27]. The Row sample and feature sample is considered with replacement to feed each decision tree and find out the result of each tree by bootstrap techniques so that the result of each decision tree is obtained and used majority voting for making the final analysis in case of regression problem mean and medium is considered. Each decision tree will have low bias and high variance but when combining them low variance will be achieved. The hyper-parameter is used to select the number of features to be considered.

3.4.5 Naive Bayes Algorithm

This type of classifier used probability techniques in which all the features have to be considered as independent of one another to calculates the probability of the values of each variable that correspond to the independent variables based on the train data [28].

Let Z be the target variable and q an array of independent variables, the probability can be obtained.

$$\mathrm{P}(\frac{Z}{q}) = \frac{\mathrm{P}(\frac{q}{F}) * \mathrm{P(Z)}}{P(q)} \tag{3}$$

3.4.5 Decision Tree Algorithm

Decision tree classifier is one of the famous techniques. It is used in a different area of research, in which the case dataset has to be divided into different branches based on the conditions. It uses a tree-look structure whereby each node in the tree is the decision point and the leaf of the node represents the output [29].

4.0 MODEL PERFORMANCE EVALUATION

Different aspects are considered before making conclusions on which Algorithm performs better than others by evaluating its accuracy. Receiver operating characteristics and area under the curve are among the components to measure in making the final assessment of the model.

4.1 Definitions of some terms used for evaluations

True Positive (TP) - Represent the positive target variables that are predicted as positive correctly.

False Positives (FP) – Represent the 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

4.2 Accuracy: Is used to determine the performance of the Algorithms

$$\text{Accuracy} = \frac{\text{Number of correct predictions made}}{\text{Total number of predictions made}} \tag{4}$$

4.3 Sensitivity: Sensitivity or recall is used to assess the number of variables that are missing in the predictions which are positive target.

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{5}$$

4.4 **Precision:** It is used to test the correctness of the model when it gives a positive outcome

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+PF}} \tag{6}$$

4.5 **Error Rate:** Evaluates the output instances that are classified incorrect.

$$\mathbf{ER} = \frac{\mathrm{FN+FP}}{TP+FP+FN+TN} \tag{7}$$

4.6 **True positive Rate (**TPR) $= \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ (8)

4.7 False Positive Rate (FPR) = $\frac{FP}{FP+TN}$ (9)

4.8 **F1 Score** = 2 * $\frac{Precision*Recall}{Precision+Recall}$ (10

5.0 RESULT AND DISCUSSIONS

5.1 Classification Models Performance

Performances of different Algorithms were checked and contrast has been made to identify the degree and accuracy of the malaria prediction algorithm that will allow the possible identification of malaria outbreaks in a particular society. Performance is measure by considering all the components futures like accuracy, Area under Curve (AUC), precision, sensitivity/Recall, specificity, F! score which is the harmonic mean of precisions and recalls of the algorithm, the rate of the error (ER), the True Positive Rate (TPR), the False Positive Rate (FPR), Average Macro and Average Weighted generated from the dataset. The overall results obtained are summarized and organized in the tables below.

Table 1. Algorithms classification performance matrix

Algorithms	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1 score (%)	Specificity (%)	Error Rate (%)	TPR (%)	FPR (%)	Avg. Macro & Weighted
SVM	93.18%	93%	91%	95%	93%	95.23%	6.82%	91.30%	4.76%	93%
KNN	86.36%	86%	81%	95%	88%	94.44%	13.62%	80.77%	5.55%	87%
Random Forest	97.72%	98%	100%	95%	98%	95.65%	2.27%	100%	4.35%	98%
Decision Tree	95.45%	95%	100%	91%	95%	91.66%	4.55%	100%	8.33%	95%
Naïve Bayes	88.63%	89%	84%	95%	89%	94.73%	11.36%	84%	5.26%	89%
Logistic Regression	95.45%	95%	95%	95%	95%	95.45%	4.52%	95.45%	4.55%	95%

Table 2. Algorithms classification performance ranking

Algorithms	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1 score (%)	Specificity (%)	Error Rate (%)	TPR (%)	TPR (%)	Average Macro & Weighted
Random Forest	1	1	1	1	1	1	1	1	1	1
Logistic Regression	2	2	3	2	2	2	2	3	2	3
Decision Tree	3	3	2	6	3	6	3	2	6	2
SVM	4	4	4	3	4	3	4	4	3	4
Naïve Bayes	5	5	5	4	5	4	5	5	4	5
KNN	6	6	6	5	6	5	6	6	5	6

Table 3. Algorithms overall performance ranking

ALGORITHM	COMBINED RANK	PERFORMANCE ORDER
Random Forest	10	1
Logistic Regression	23	2
Decision Tree	36	3
Support Vector Machine	37	4
Naïve Bayes	47	5
K-Nearest Neighbour	57	6

Table 1 above, shows the percentages and performances of each classifier which indicate the significance of one classifier over another, Random forest has the highest level of performance in accuracy more than all other classifiers which are involved in the process of accessing the ability of every classifier to produce a result with higher precision and AUC for the prediction of malaria outbreak base on the variables involves in the analysis.

From table 2, It indicates the ranking performances in which the one with higher performance is assigned to a numerical value one followed by the second with which carry the numerical value of two continuously up to the last with numerical value six. For Recall/ Sensitivity measures all the classifiers have the same value except for the Decision Tree classifier.

After adding each row from table 2, the classifier with less number is considered the best classifier. The performance of the Random Forest Algorithm is the best with (97.72%) accuracy, (98%) AUC, and (100%) precision followed by logistic Regression. In the same vain equally analysis has indicated classifiers such as KNN and Naïve Bayes with a law ability to detect and predict the malaria outbreak in a given society base on the availability of the response variables and in turn, it has a smaller percentage in term of performances. Some classifier has demonstrated their average performance toward prediction power and ability to detect malaria outbreak such as Decision tree and SVM.

The best three performing Classifiers in our use case are found to be Random Forest, Logistic Regression, and Decision Tree Algorithm while Naïve Bayes have less performance.

5.2 C**ONCLUSION**

The results signified that atmospheric factors, non-climatic features are significant in determining the occurrence of malaria outbreaks in a given society. Also implies one Algorithm outperforms better than others in given a proper and accurate result based on the dataset used.

Accurate prediction is essential in determining when, where, and how the outbreak will occur so that people will get to know how to prepare for it and take all the necessary measures to ensure their safety against the causative agent of death.

Since malaria can be prevented, if a certain measure is taking into account then Government and non-Governmental organizations have to use the available information in providing all the necessary resources that can help individuals in fighting against malaria more especially in the tropical and temperate regions of sub-Sahara in Africa. Programs like Global Technical Strategy for Malaria 2016-2030 [30]; have to encourage and public campaigns on the effect of malaria and how peoples can manage their environmental sanitation effectively.

5.3 FUTURE WORK

This research can be improved by using hybridized ensemble Method of machine learning to realize a productive performance model by integrating atmospheric features and non-climatic features which include population size, the vegetation of the area, nature of interventions received, and other environmental features together during prediction. It's highly recommended to enlighten the public on the effect of malaria and publicized its monthly data of each country or region as it happens with the covid-19 pandemic. So that more datasets about malaria will be available at any time because most of the current malaria datasets tend to be incomplete and vague.

REFERENCES

1. World Health Organization. Malaria Rapid Diagnostic Test Performance: Results of WHO Product Testing of Malaria RDTs: Round 6; World Health Organization: Geneva, Switzerland, 2015.
2. World Health Organization. Malaria Report 2020.
3. Roll Back Malaria, “Climate change and Malaria,” 2015, http:// www.google.com.gh/search?q=climate+change+and+malaria+ 2015.pdf&client=ms-opera-mini-android&channel=new&gws rd=cr&ei=tUzUVrytJ8uIaMHmiMAF.
4. Haque, U.; Hashizume, M.; Glass, G.E.; Dewan, A.M.; Overgaard, H.J.; Yamamoto, T. The role of climate variability in the spread of malaria in Bangladeshi highlands. PLoS ONE 2010, 5, e14341.
5. Abeku TA. Response to malaria epidemics in Africa. Emerg Infect Dis. 2007; 13:681–6. 4.
6. Maes P, Harries AD, Van den Bergh R, Noor A, Snow RW, Tayler-Smith K, et al. Can timely vector control interventions triggered by atypical environmental conditions prevent malaria epidemics? A case-study from Wajir County, Kenya. PLoS ONE. 2014;9: e92386. 5.
7. Checchi F, Cox J, Balkan S, Tamrat A, Priotto G, Alberti KP, et al. Malaria epidemics and interventions, Kenya, Burundi, Southern Sudan, and Ethiopia, 1999–2004. Emerge Infect Dis. 2006; 12:1477–85.
8. M. Woube, “Geographical distribution and dramatic increases in incidences of malaria: consequences of the resettlement scheme in Gambela, SW Ethiopia,” Indian Journal of Malariology, vol. 34, no. 3, pp. 140–163, 1997.
9. T. A. Abeku, G. J. van Oortmarssen, G. Borsboom, S. J. de Vlas, and J. D. F. Habbema, “Spatial and temporal variations of malaria epidemic risk in Ethiopia: factors involved and implications,” Acta Tropica, vol. 87, no. 3, pp. 331–340, 2003.
10. H. D. Teklehaimanot, M. Lipsitch, A. Teklehaimanot, and J. Schwartz, “Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms,” Malaria Journal, vol. 3, article 41, 2004.
11. A. D. Kassa and B. B. Beyene, “Climate variability and malaria transmission—fogera district, Ethiopia, 2003–2011,” Science Journal of Public Health, vol. 2, no. 3, pp. 234–237, 2014.
12. M. C. Thomson, S. J. Mason, T. Phindela, and S. J. Connor, “Use of rainfall and sea surface temperature monitoring for malaria early warning in Botswana,” The American Journal of Tropical Medicine and Hygiene, vol. 73, no. 1, pp. 214–221, 2005.
13. O. Ndiaye, J.-Y. le Hesran, J.-F. Etard et al., “Variations climatiques et mortalite attribu ´ ee au paludisme dans la zone de ´ niakhar, sen´ egal de 1984 ´ A 1996,” ` Sante´, vol. 11, pp. 25–28, 2001.
14. C. Christiansen-Jucht, P. E. Parham, A. Saddler, J. C. Koella, and M.-G. Basa´nez, “Temperature during larval development ˜ and adult maintenance influences the survival of Anopheles gambiae s.s,” Parasites & vectors, vol. 7, 2014.
15. Y. A. Afrane, A. K. Githeko, and G. Yan, “The ecology of Anopheles mosquitoes under climate change: case studies from the effects of deforestation in East African highlands,” Annals of the New York Academy of Sciences, vol. 1249, no. 1, pp. 204–210, 2012.
16. 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)
17. 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.
18. 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
19. Caravaca Moreno, J., Soria Olivas, E., Bataller Mompeán, M., Serrano López, A. J., Such Miquel, L., Vila Francés, J., & Guerrero Martínez, J. F. (2014). Application of machine learning techniques to analyse the effects of physical exercise in ventricular fibrillation. Computers in Biology and Medicine, 2014, vol. 45, num. 1, p. 1-7.

20. 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.
21. 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.
22. 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)
23. Engineering (IJCSEE) ,Volume 1, Issue 1 (2013) [10] http://machinelearningmastery.com/a-tour-of-machine-learning algorithm
24. S. Patel, 'Chapter 2 SVM (Support Vector Machine) - Theory', Machine Learning 101, [Online], Available: https://medium.com/machine-learning-101/chapter-2-svm-support- vector-machine-theory-f0812effc72. [Accessed: 28 - Sep – 20.
25. T. Srivastava, 'Introduction to k-Nearest Neighbors: Simplified (with implementation in Python)', Analytics Vidhya. 2018 [Online]. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k- neighbours-algorithm- -clustering/. [Accessed: 19 - Sep – 2018.
26. Difference Between Linear and Logistic Regression', Tech Differences, [Online], Available: https://techdifferences.com/difference-between-linear-and-logisticregression.html. [Accessed:
27. N. Donges, 'The Random Forest Algorithm', Towards Data Science. 2018 [Online]. Available: https://towardsdatascience.com/the-randomforest-algorithm-d457d499ffcd. [Accessed: 19 - Sep - 2018].
28. S. Patel, 'Supervised Learning and Naive Bayes Classification', Machine Learning 101, [Online], Available: https://medium.com/machine-learning-101/chapter-1-supervisedlearning-and-naive-bayes-classification-part-1-theory-8b9e361897d5. [Accessed: 28 - Sep - 2018].
29. Machine Learning Algorithms List [2021 Updated] (simplilearn.com)
30. WHO, 'Global technical strategy for malaria 2016-2030', 2015.