UNIVERSITÉ DU BURUNDI
GRENIER DU SAVOIR DU BURUNDI
Dépôt institutionnel officiel
https://repository.ub.edu.bi

Exploring predictive frameworks for malaria in Burundi

Mfisimana, Lionel Divin; Et al.

2022-03
Infectious Disease Modelling 7
https://repository.ub.edu.bi/handle/123456789/2151

Document téléchargé depuis le **Grenier du Savoir du Burundi** — https://repository.ub.edu.bi

Infectious Disease Modelling 7 (2022) 33–44

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Exploring predictive frameworks for malaria in Burundi

Lionel Divin Mfisimana [a, 1], Emile Nibayisabe [a, 1], Kingsley Badu [b], David Niyukuri [a, c, d, *]

[a] Faculté des Sciences Fondamentales, Institut Supérieur des Cadres Militaires, Burundi
[b] Department of Theoretical and Applied Biology, Kwame Nkrumah University of Science and Technology, Ghana
[c] Division of Epidemiology & Biostatistics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[d] The South African Department of Science and Technology–National Research Foundation (DST-NRF) Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, Cape Town, South Africa

ARTICLE INFO

Article history:
Received 13 July 2021
Received in revised form 22 February 2022
Accepted 5 March 2022
Available online 9 March 2022
Handling Editor: Dr Lou Yijun

Keywords:
Burundi
Malaria
Modelling
Generalized linear model
Neural network

ABSTRACT

In Burundi, malaria infection has been increasing in the last decade despite efforts to increase access to health services, and several intervention programs. The use of heterogeneous data can help to build predictive models of malaria cases. We built predictive frameworks: the generalized linear model (GLM), and artificial neural network (ANN), to predict malaria cases in four sub-groups and the overall general population. Descriptive results showed that more than half of malaria infections are observed in pregnant women and children under 5 years, with high burden to children between 12 and 59 months. Modelling results showed that, ANN model performed better in predicting total cases compared to GLM. Both model frameworks showed that education rates and Insecticide Treated Bed Nets (ITNs) had decreasing effects on malaria cases, some other variables had an increasing effect. Thus, malaria control and prevention interventions program are encouraged to understand those variables, and take appropriate measures such as providing ITNs, sensitization in schools and the communities, starting within high dense communities, among others. Early prediction of cases can provide timely information needed to be proactive for intervention strategies, and it can help to mitigate the epidemics and reduce its impact on populations and the economy.

© 2022 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Malaria infection has been, and still a major public health concern in Burundi (World malaria report 2017, 2018). The World Health Organization (WHO) estimates show that more than 80% of the Burundian population are at high risk of acquiring malaria infection (World malaria report 2017, 2018). Malaria cases increased from 2.6 million in 2013 to 8.3 million in 2016 with a further 18 percent increase in the first half of 2017 compared to 2016 (Malaria Initiative, 2016). In 2017 malaria was declared a national epidemic (Burundi Humanitari, 2017; Lok & Dijk, 2019; Weekly Bulletin on Outbre, 2017). According to the

* Corresponding author. Faculté des Sciences Fondamentales, Institut Supérieur des Cadres Militaires, Burundi.
E-mail addresses: lili.divin8@gmail.com (L.D. Mfisimana), enibayisabe472@gmail.com (E. Nibayisabe), kingsbadu@gmail.com (K. Badu), kuriniyu@gmail.com (D. Niyukuri).

Peer review under responsibility of KeAi Communications Co., Ltd.

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.idm.2022.03.003
2468-0427/© 2022 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Institute for Health Metrics and Evaluation (IHME), in 2019 malaria was ranked fourth cause of death after Diarrheal diseases, Neonatal disorders, and tuberculosis in Burundi (Health metrics for Burund, 2020).

The health system still has many challenges, despite efforts by the Government to improve by increasing access to healthcare facilities and technical capacities (Sinzinkayo et al., 2021). Burundi suffers from a shortage of qualified personnel, medical resources, and high burden of diseases (Lozano & Garrido, 2015). According to the World Bank statistics in 2017, Burundi was estimated to have 0.1001 physicians for 1000 people (World health organization, 2017), and in 2010, the density of doctors, nurses and midwives combined was estimated to be 2 per 10,000 population (Density of doctors and nurse, 2010).

In a study published in 2011 by Nkurunziza et al. (Nkurunziza et al., 2011), the authors claimed that 17.4% of patients did not have access to health care at all, while 81.5% of patients were forced to go into debt or sell their property to pay for the health costs. Furthermore, the authors of same study (Nkurunziza et al., 2011), and the report of health systems (Lozano & Garrido, 2015) reported a huge disparity between the urban areas especially the economic capital Bujumbura, and the other parts of the country.

According to the World Vision International, an international Non Government Organization (NGO) which support malaria interventions in Burundi, some major drivers of malaria infection in Burundi are: the low usage of Insecticide Treated Bed Nets (ITNs), climate change, population density, shifting of farming practices, food insecurity, and lack of knowledge and action to prevent malaria in the communities (Eight facts about burundi, 2017).

Studies have shown association of malaria cases with temperature, humidity, rainfall, usage of ITNs, geographical location, and socioeconomic factors (Beck-Johnson et al., 2013; Li et al., 2013; Pascual et al., 2006; Semakula et al., 2015). This means that the data of these variables can be used to predict number of malaria cases at a certain level of certainty. In Uganda it was reported that malaria is highly endemic with temperatures and precipitation that allow stable transmission throughout the year with relatively low seasonal variability in most parts of the country (Siya et al., 2020). In areas with a long dry season such as the Sahel, and particularly in Burkina Faso, rains determine the abundance of mosquito populations and thus the intensity of transmission with mosquitoes then benefiting from multiple puddles to reproduce and also high humidity levels of the surrounding air. In Niger, a 16% increase in rainfall between 2005 and 2006 was accompanied by a 132% increase in mosquito abundance in the village of Banizoumbou, located in the Sahelian zone (Bomblies et al., 2008). The impact of altitude on malaria transmission is directly related to the decrease in temperature that influences both the vector (anopheles mosquito) and the parasite (Bayoh & Lindsay, 2003; Kipruto et al., 2017). We know also that the wind speed varies according to the altitude so that at high altitude mosquitoes may lose their vectorial capacity.

Using multiple data sources for malaria prediction has been argued to complement the surveillance system of malaria (Wang et al., 2016) which is often based on reported data. Therefore, given the current knowledge on factors which may increase or decrease malaria prevalence, it is important to enhance decision making, and resource allocation with targeted interventions for effective malaria control. And since there is no one-size fits all intervention to eliminate malaria, modelling frameworks can help to explore a range of interventions given different predictions of infection cases, epidemic estimates, and cost effectiveness. This is crucial for decision making to be able to tailor interventions and optimize resources as stated in the WHO's Global Technical Strategy for Malaria 2016–2030 (Global technical strategy, 2021).

The objective of this study was to use available data sources which have been proven to be associated to malaria infection (Haque et al., 2011; Onyango et al., 2016; Semakula et al., 2015; Thang et al., 2008) to build basic predictive frameworks to predict malaria cases in different sub-groups in Burundi, namely response variables: pregnant women and children under 5 years, among pregnant women, children between 0 and 11 months, children between 12 and 59 months, and the overall general population. From those models outputs, we should identify predictors variables which more important for the increase and decrease of malaria cases, and also get to know their magnitude. This provides the starting point for more complex models with granular data to understand different factors influencing the transmission dynamics of malaria, and subsequent control measures in Burundi.

2. Methods

2.1. Data

The study was conducted using monthly data collected from different provinces of Burundi for the period 2010–2017. The data consists of records of total malaria cases, cases among children under 5 years and pregnant women, as well as the number of insecticide treated nets (ITNs) distributed to children and pregnant women. We also used monthly climate data, and demographics data including annual population estimates and schooling rates for all provinces.

A record of the total nationwide malaria cases was available from the National Integrated Malaria Control Program of the Ministry of Health. Detailed monthly data sets on malaria infection for children (from 0 to 11 months then up 12–59 months male and female), and pregnant women (before and after first trimester of pregnancy), the number of ITNs distributed (for children under five years and pregnant women) per each province were available from the Department of National System of Health Information of the Ministry of Health.

Monthly climate data (weather) per province were provided by the Geographical Institute of Burundi. In addition, demographic data for each province were provided by the open library of the National Institute of Statistics. And the schooling rates data per province per year were provided by the Planning Office of the Ministry of Education and Scientific Research.

Thus, data were collected on different time scales since some data sources were monthly while others yearly. All malaria epidemic data (malaria cases in children girls and boys between 0 and 11 months, 12 and 59 months, pregnant women, total, number of pregnant women, number of children below 5 years, total ITNs distributed to women and children) were collected at monthly basis in provinces. Climate related data (rainfall, temperature, humidity) were also collected at monthly basis in different stations in provinces. But demographics (populations, education, and location) data were collected once a year except for location (longitude, latitude, altitude, and area). After cleaning, handling missing values as discussed in the following subsection 2.2, and correlation analysis, in our modelling exercise, among 9 variables we used, 3 of them (education rates, population density and altitude) were not captured at monthly basis in our data. Those collected once a year (population density and education rates) were assumed to be the same along all month of the year and change with new data of the year, and the altitude which is constant, it was constant all months and years.

2.2. Data imputation

We combined all the databases in one data frame, after cleaning, we had a new data frame with the following variables: area, longitude, latitude, altitude, rainfall, temperature, humidity, population, education, malaria cases among children under 11 months, malaria cases among children between 12 and 59 months, malaria cases among pregnant women with less than 3 months, malaria cases among pregnant women beyond 3 months, severe malaria cases, ITNs given to children, ITNs given to pregnant women, severe malaria cases, and other malaria cases. From climate data, we have data on temperature, humidity, and rainfall. From demographics, we have population, education rates, and geographical location (longitude, altitude, latitude, and surface area).

The climate data had missing values, and to handle them we used a non-parametric missing value imputation approach using random forest method (missForest) (Stekhoven, 2015), by specifying the maximum number of iterations to be performed (100) and the number of trees to grow in each forest to be 1000. This method is often used to impute continuous and/or categorical data including complex interactions and nonlinear relations. This was done with the data frame of combined data from malaria infection, climate, and demographics.

With same combined data frame, we performed a correlation analysis. Before the analysis, we had to drop some variables from malaria epidemiological data (severe cases, and other subgroups such as malaria cases among different periods of pregnancy for women), and we kept only sub-groups of interest in our modelling exercise (total malaria cases, and cases among children between 0 and 11 months, between 12 and 59 months, below 5 years, pregnant women, and pregnant women and children below 5 years). We also drop the latitude variable from the location, since we assumed that longitude and altitude were sufficient given the limited size of country.

The results of the correlation analysis showed a triangle like clusters of highly correlated variables as we can see at Fig. 1: the first one was among malaria cases in different sub-groups; the second one was between the number of pregnant women, number of children below 5 years, and the number of ITNs distributed in the general population; and the third one was between malaria cases in different sub-groups and the number of pregnant women, number of children below 5 years, and the number of ITNs distributed. We also had other non negligible correlations between the area and the number of pregnant women and children under five years, and small but positive correlation with the population, thus, that lead us to create a new variable, the population density. We used it instead of using population, and area. The well known negative correlation between temperature and altitude (Lancaster, 1980; Peng et al., 2020) was also observed. Longitude was much positively correlated to sub-groups of malaria cases, surface area and some populations (number of women and children under five years), and since we had to keep a priori sub-groups of malaria cases, and had created a new variable with population and area, we dropped the longitude.

Based on correlation analysis results, to build the predictive models for malaria cases, we used the following variables: altitude, rainfall, temperature, humidity, population density (a variable we created with population and surface areas), education rates, number of pregnant women, number of children under five years, and number of ITNs distributed. We built the models using randomly 75% of the data set for training, and 25% for testing. In order to be more realistic, we also considered a scenario where we used data from 2010 up to 2016 for training, and predict 2017 cases. For that scenario we considered only predicting the monthly total cases of malaria.

2.3. Modelling: general linear model and artificial neural network model

We built a general linear model (GLM) (Bates et al., 2014), and an artificial neural network model (ANN) (Gurney, 1997) to predict malaria cases in different scenarios. Each response variable (malaria cases among pregnant women and children under 5 years, malaria cases among pregnant women, malaria cases among children between 0 and 11 months, malaria cases among children 12 and 59 months, and overall total malaria cases) was evaluated by both types of models. To ensure model validity and performance, we used cross-validation approach with 100 fold replication. To assess the goodness of fit, we computed the mean square root error (MSRE) of the difference between true and predicted values of malaria cases.

The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allow the dependent variables to be non-normal (McCullagh & Nelder, 2019). Denote X as the independent variables and Y the dependent variables. Then the expected value of Y conditional on X is given by

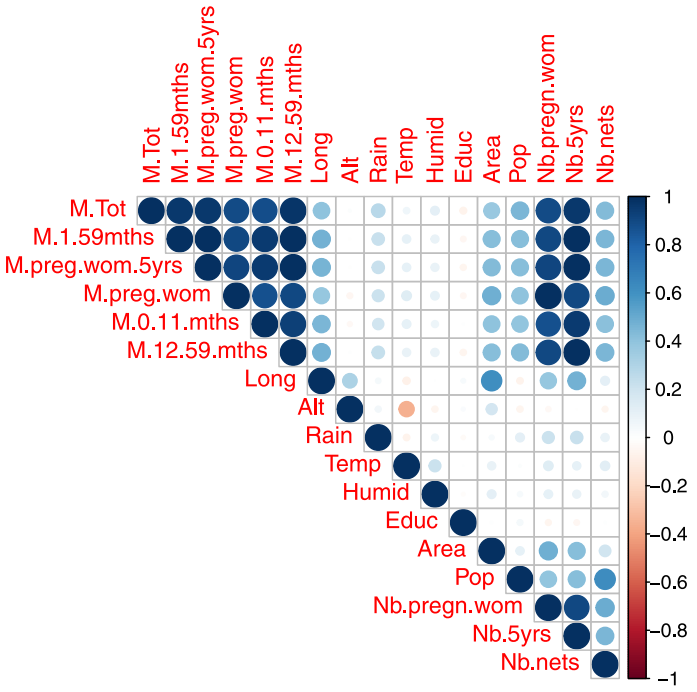


Fig. 1. Correlation matrix between total malaria cases (M.tot), malaria cases among children between 0 and 11 months (M.0.11.mths), malaria cases among children between 12 and 59 months (M.12.59.mths), malaria cases among children under five years (M.1.59.mths), malaria among pregnant women (M.preg.wom), malaria among pregnant women and children under five years (M.preg.wom.5yrs), population (Pop), number of pregnant women (Nb.pregn.wom), number of children under five years (Nb.5yrs), number of bed net distributed (Nd.nets), longitude (Long), altitude (Alt), temperature (Temp), rainfall (Rain), humidity (Humid), education rates (Educ), and surface area (Area).

$$E(Y|X) = g^{-1}(X\beta) \tag{1}$$

where g is called the link function, and β are linear combination parameters. A GLM model consists of three components: (i) random component, specifying the conditional distribution of the response variable (Y), (ii) linear predictor—that is a linear function of regressors ($X\beta$), and (iii) a smooth and invertible linearizing link function $g()$, which transforms the expectation of the response variable, $\mu = E(Y)$, to the linear predictor. In our GLM model, the link function was a Gaussian distribution. If we assume the linear predictor to be μ, we will have

$$\mu = \sum_{i=1}^{9} \beta_i x_i \tag{2}$$

where x_i are the nine predictor variables (population density, number of pregnant women, number of children below 5 years, number of ITNs, education rates, humidity, rainfall, altitude, and temperature).

The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. With GLM fitting to data, we get effect of each predictor to the response (output) with assumption that other predictors are omitted. To be able to see the effects of all the predictors we performed an ANOVA to the GLM outputs with Chi-Square test which consider adding sequentially terms (predictors) to the model (McCullagh & Nelder, 2019).

Artificial neural networks are tools used in machine learning which are able to detect multiple nonlinear interactions among a series of input variables (predictors). They are brain-inspired systems which are intended to replicate the way that humans learn. Neural networks consist of input and output layers, as well as a hidden layer consisting of units that transform the inputs into a stable form for the output layer, and each layer has neurons (nodes or vertices). The number of nodes of the input layer is equal to the number of inputs variables, and the number of nodes of the output layer is equal to the number of outputs. For our model, the number of the nodes of the input layer was nine and for the output layer was one. To build the

hidden layers, we followed the rule of thumbs which describes how to choose the number of hidden layers and nodes in a feedforward neural network (Heaton, 2008). Similarly to GLM model above, if y denotes the output which was malaria cases in the five sub-groups, and x_i are the nine predictor variables, the symbolic description of the ANN model was given by the following expression.

$$y \sim \sum_{i=1}^{9} \beta_i x_i \tag{3}$$

After exploration of different artificial neural network configurations with number of layers and neurons, we considered an artificial neural network with two hidden layers with respective internal nodes (3, 4), and the neural network was trained 10 times using *neuralnet* package (Günther & Fritsch, 2010). For the ANN model, we had to normalize the data for neural network model using the minimum and maximum values (Gökhan et al., 2019). For the ANN models, we computed the relative variable importance using connection weights approach by Olden et al. (Olden et al., 2004) which is implemented in NeuralNetTools R package (Beck, 2022). The results showed the importance of predictors for any given outcome (malaria cases sub-groups) and the direction of that importance.

For any of the above mentioned modeling approaches, the goodness of fit was measured by the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x_i^*)^2} \tag{4}$$

where x_i is the observed number of cases and x_i^* the predicted number of cases. To evaluate the models' performance, we used cross-validation (CV) technique, we re-sampled 10 times the data and run the models, and we summarized the models' performance by the mean of RMSE values which were computed per each time we ran a model. All the computations were performed using R (version 3.4.4) software (R Core Team, 2018) in a Linux environment.

3. Results

3.1. Descriptive analysis

A descriptive analysis of annual malaria cases shows that the epidemic curves for the five categories between 2010 and 2017 have an increasing trend as we can see at Fig. 2. The epidemic curves show what is well known that pregnant women and children under 5 years are at risk for malaria infection. However, although there are a lot of malaria cases for pregnant women and children under 5 years, almost a half of cases came from other populations groups of adults and children above 5 years.

When this high risk group is divided in three subgroups with pregnant women, children between 0 and 11 months, children between 12 and 59 months, we can see that children between 12 and 59 months have a disproportionate burden of malaria infection compared to other subgroups in all the eight years (2010–2017), with pregnant women group being the one

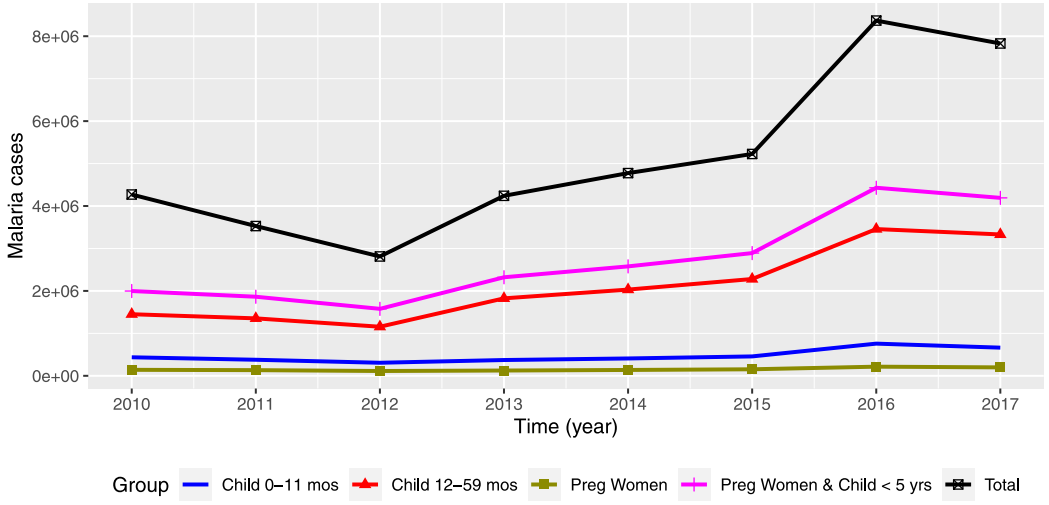


Fig. 2. Country level epidemic curves of malaria cases in children between 0 and 11 months, children between 12 and 59 months, pregnant women, pregnant women and children under 5 years, and the overall total cases between 2010 and 2017.

with low cases. Regarding geographical distribution, the charts of province level cases at Fig. 3 show that provinces of Kirundo, Gitega, Muyinga, and Ngozi were much affected compared to others.

When we aggregate the monthly number of malaria cases for the eight years (2010–2017), Fig. 4 shows that the total malaria cases have same trend as cases among pregnant women and children under 5 years, and cases among children between 12 and 59 months. We can say that the temporal trend of malaria cases among pregnant women and children under 5 years can benchmark the malaria epidemic in Burundi. However, children between 12 and 59 months have same trend as cases pregnant women and children under 5 years, but this is explained by the fact that the large proportion of the latter category is made of children between 12 and 59 months.

As we indicated earlier, in the descriptive subsection 3.1, in the 8 years period (2010–2017), the epidemic curves of malaria cases has an increasing trend. Considering malaria cases in pregnant women and children under 5 years as a benchmark of the trend (from Fig. 4), Fig. 5 shows that, even the year with the highest number of cases (2016), it did not have high cases in all months. A visual comparison between monthly cases among pregnant women and children under 5 years in the 8 years (2010–2017) showed that there has been disparities (Fig. 5). If we compare the monthly number of cases in those 8 years from Fig. 5, it shows that although the magnitude is different with much disparities within months, the overall trend across months of the year is similar of what we see at Fig. 4. In January, we have high number of infections which decreases in February and start increasing by March up to June after which a sharp decrease is observed in summer with lowest cases in August. In September, cases start increasing again up to December.

3.2. Models outputs

The results of the GLM models built for malaria cases (total cases, among children under 5 years and pregnant women, among pregnant women, and among children between 0 and 11 months, and those between 12 and 59 months) are given in Table 1. Among the five sub-groups of malaria cases, the overall total number of malaria cases was the only response variable where all parameters estimates associated to predictors have statistically significant (based on their p-values, $Pr(>|t|)$ in Table 1) effect on the outcome, followed by pregnant women which has 6 among the 9 predictors. Other responses variables had limited number of statistically significant predictors.

We can see in Table 1 that some of the variables have an increasing or decreasing effect on malaria cases. Rainfall, humidity, number of pregnant women, and the number of children under five years have an increasing effect on the overall malaria cases. But, education, temperature, and the number of ITNs distributed to pregnant women and children have a decreasing

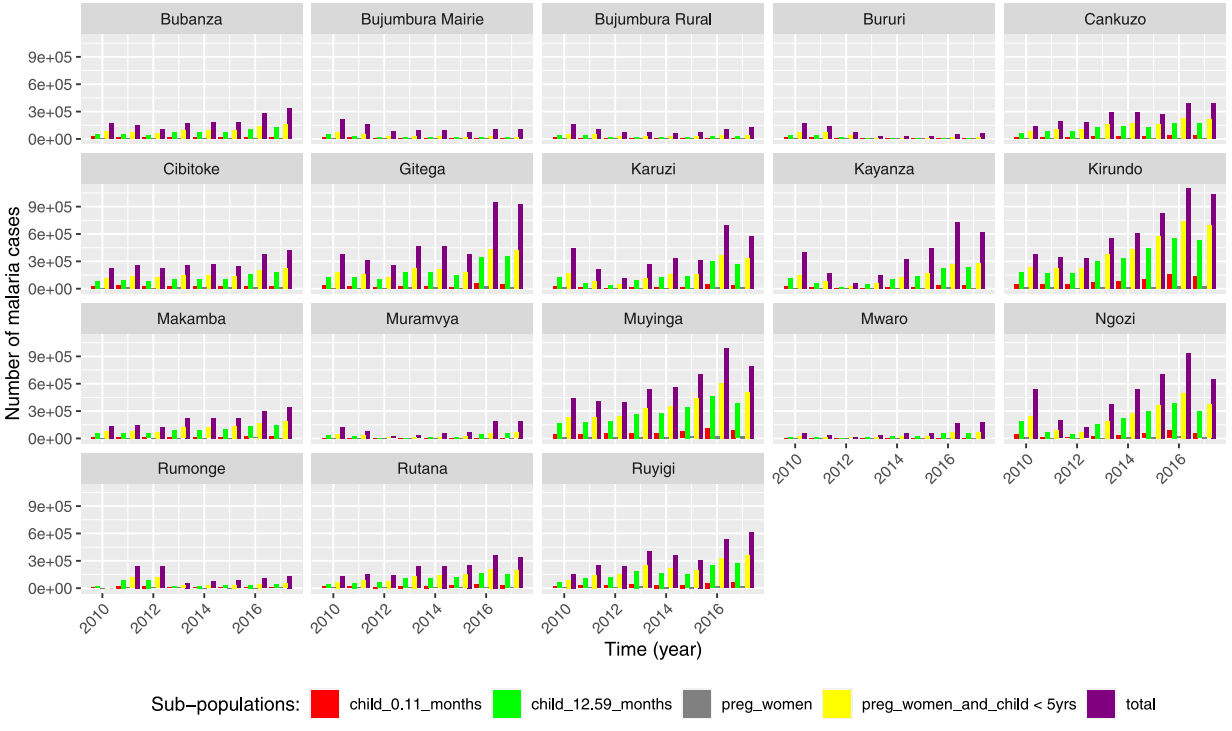


Fig. 3. Province level malaria cases in children between 0 and 11 months, children between 12 and 59 months, pregnant women, pregnant women and children under 5 years, and the overall total cases between 2010 and 2017.

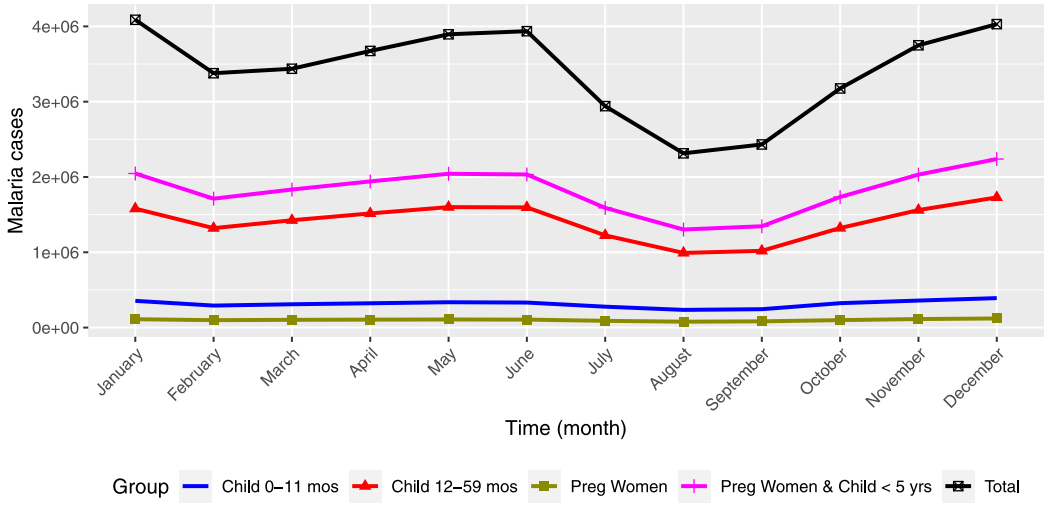


Fig. 4. Aggregating monthly malaria cases in different sub-groups between 2010 and 2017.

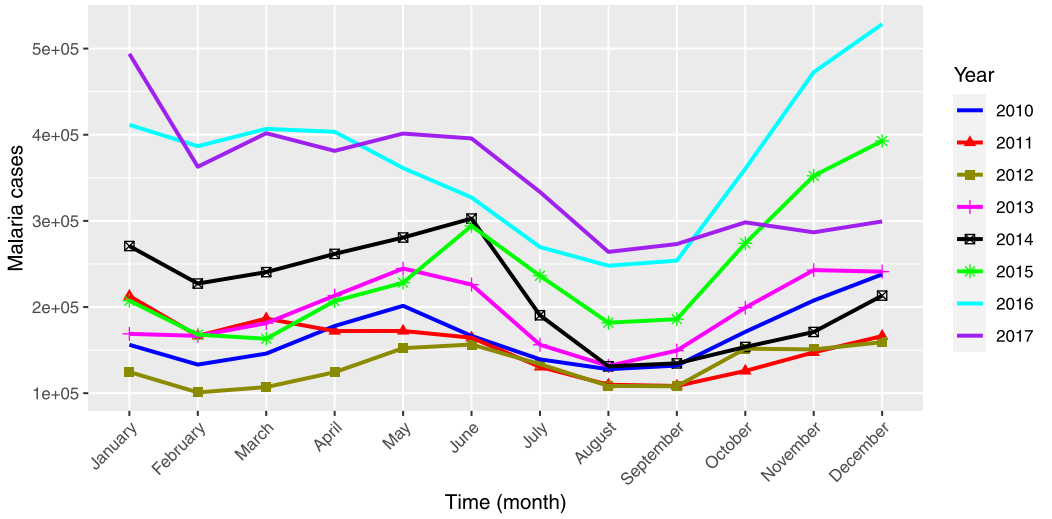


Fig. 5. Monthly malaria cases in pregnant women and children under 5 years in 8 years time period (2010–2017).

effect on the overall malaria cases. In addition, by sequentially adding one predictive variable for all the generalized models with results in Table 1, the contribution of each variable was significant as we can see for columns of the statistical significance of the Chi-square test ($Pr(>Chi)$), except for ITNs variables in predicting malaria cases among pregnant women and children below 5 years.

From the ANN model, the computed relative variables' importance in the five scenarios of different sub-groups of malaria cases (see Fig. 6) showed that for the overall malaria cases, rainfall, population density and the number of children under 5 years were much predominant positive predictors whereas education and ITNs were relatively dominant negative predictors.

The Root Mean Square Error (RMSE) and Cross Validation (CV) between observed and predicted values by the GLM and ANN models for malaria cases in the overall population (total cases), among pregnant women and children under 5 years, among pregnant women, among children between 0 and 11 months, and those between 12 and 59 months are given in Table 2. The error values show that ANN performs better in estimating total malaria cases compared to GLM model (Table 2). The generalized linear models do not perform too bad since the difference between generalized linear model and neural network model is between two and three thousands cases (CV and MSRE). For cases among pregnant women, and children, the generalized linear model performed better compared to neural network model (Table 2).

Table 1
Parameter estimates from the generalized linear models (GLM) for malaria cases, and the significance level of the results of the ANOVA performed to the GLM outputs with Chi-Square test.

Parameters	Total			Pregnant women and children (<5 years)			Pregnant women			Children (0–11 months)			Children (12–59 months)		
	Estimate	*Pr*(> |*t*|)	*Pr*(> *Chi*)	*Estimate*	*Pr*(> |*t*|)	*Pr*(> *Chi*)	*Estimate*	*Pr*(> |*t*|)	*Pr*(> *Chi*)	*Estimate*	*Pr*(> |*t*|)	*Pr*(> *Chi*)	*Estimate*	*Pr*(> |*t*|)	*Pr*(> *Chi*)
Intercept	1009.211 (−1650, 3669)	0.457	NA	250.498 (91, 409)	0.002	NA	22.02 (9, 34)	<0.001	NA	192.184 (−164, 549)	0.292	NA	5.705 (−358, 370)	0.976	NA
Altitude	1.187 (0.367, 2.008)	0.005	0.6753	−0.001 (−0.052, 0.05)	0.97	0.003	−0.016 (−0.02, −0.012)	<0.001	<0.001	−0.139 (−0.25, −0.028)	0.015	<0.001	0.073 (0.043, 0.189)	0.215	0.244
Rainfall	15.965 (13.359, 18.57)	<0.001	<0.001	0.023 (−0.046, 0.092)	0.517	<0.001	0 (−0.005, 0.005)	0.996	<0.001	−0.488 (−0.834, −0.142)	0.006	<0.001	0.287 (0.131, 0.443)	<0.001	<0.001
Temperature	−140.063 (−221.831, −58.295)	0.001	<0.001	−2.511 (−7.391, 2.369)	0.313	<0.001	0.21 (−0.161, 0.581)	0.267	<0.001	1.584 (−9.383, 12.551)	0.777	<0.001	2.42 (−8.77, 13.609)	0.672	<0.001
Humidity	21.935 (7.174, 36.696)	0.004	<0.001	−0.438 (−1.369, 0.492)	0.356	<0.001	−0.03 (−0.096, 0.035)	0.366	<0.001	−2.275 (−4.348, −0.202)	0.032	0.179	1.377 (−0.677, 3.431)	0.189	<0.001
Education	−5.615 (−8.569, −2.661)	<0.001	<0.001	−1.229 (−1.405, −1.053)	<0.001	<0.001	0.041 (0.022, 0.06)	<0.001	<0.001	1.71 (1.29, 2.13)	<0.001	<0.001	−1.69 (−2.128, −1.252)	<0.001	<0.001
Population density	0.675 (0.47, 0.881)	<0.001	<0.001	0.002 (−0.011, 0.015)	0.738	<0.001	−0.003 (−0.003, −0.002)	<0.001	<0.001	0.032 (0.003, 0.06)	0.03	<0.001	−0.036 (−0.066, −0.005)	0.021	<0.001
Pregnant women	5.648 (4.36, 6.936)	<0.001	<0.001	1.104 (1.027, 1.181)	<0.001	<0.001	0.993 (0.987, 0.998)	<0.001	<0.001	0.201 (0.027, 0.375)	0.024	<0.001	−0.193 (−0.366, −0.02)	0.029	<0.001
Children < 5years	1.444 (1.387, 1.5)	<0.001	<0.001	0.992 (0.989, 0.995)	<0.001	<0.001	0.001 (0, 0.001)	<0.001	<0.001	0.186 (0.178, 0.193)	<0.001	<0.001	0.815 (0.808, 0.823)	<0.001	<0.001
ITNs	−0.216 (−0.459, 0.027)	0.081	0.081	0.005 (−0.011, 0.02)	0.567	0.567	−0.002 (−0.003, −0.001)	0.002	0.002	−0.073 (−0.107, −0.04)	<0.001	<0.001	0.044 (0.01, 0.079)	0.013	0.0124

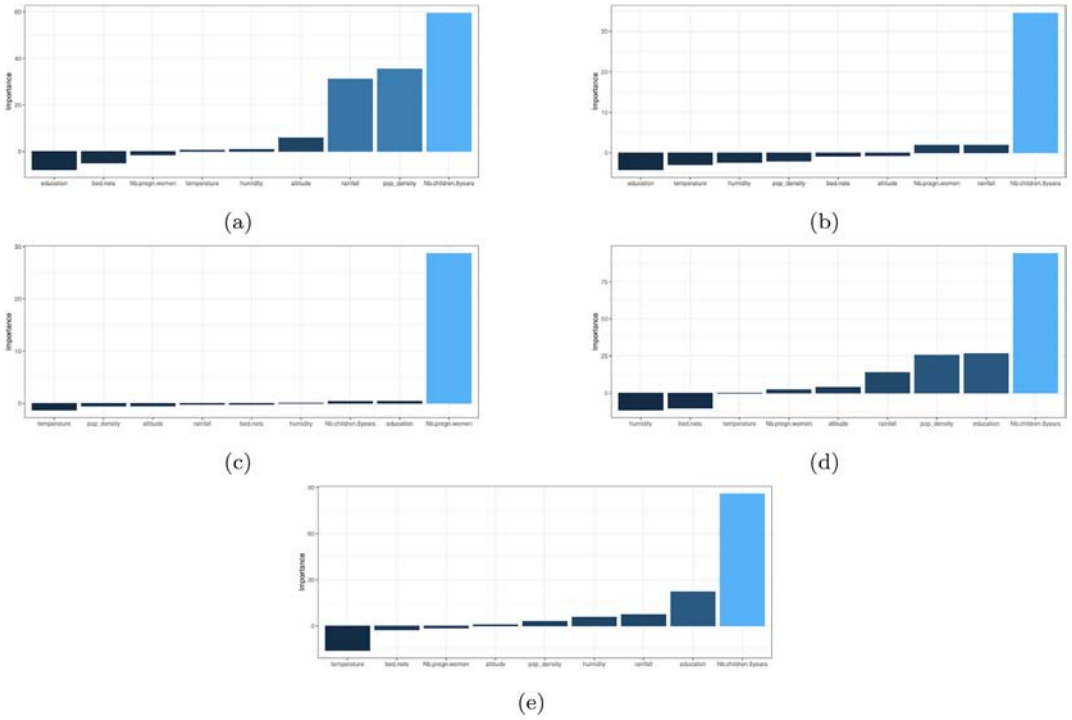


Fig. 6. Variables importance in the artificial neural network model to predict (a) overall total malaria cases, (b) cases among pregnant women and children below 5 years, (c) cases among pregnant women, (d) cases among children between 0 and 11 months, and (e) cases among children between 12 and 59 months.

The total monthly predicted malaria cases for 2017 (Fig. 7) was 7.81 millions with neural network model, and 7.48 millions with generalized linear model, but the observed data was 7.82 millions. Overall, the neural network model performed better compared to generalized linear model. Looking on months data, we can see that in some months one model would have better predictions compared to another.

4. Discussion

Splitting the population in sub-groups shed lights on more high risk groups to which interventions should be intensified. We have seen much proportion of cases were among children between 12 and 59 months. It is much likely that we may discover different transmission patterns across different age groups within other no reported groups of adults and children above 5 years. Thus, having individual level data can contribute to enhance interventions. The high burden of children between 12 and 59 months is supported by the literature which shows that new born are relatively protected for malaria infection since they can acquire immunity to malaria from their mothers who were exposed to the infection, but this immunity decrease by 6–9 months (Dobbs & Dent, 2016; Doolan et al., 2009).

It is important to note that despite higher number of malaria cases for pregnant women and children under 5 years, other sub-groups of adults and children above 5 years are also bearing non-negligible burden of malaria infection. These other groups which are supposed to have stronger immunity compared to pregnant women and children below 5 years play a key role in the transmission cycle of malaria. Therefore, specific interventions should target these groups, thus, epidemiological and behavioural studies should be conducted to better understand the main drivers and risk factors of malaria infection in these groups.

From the descriptive results, some months of the year are characterised by higher cases compared to others, and others have very low cases. This may be explained by weather changes and rains, but we have seen that the monthly trends in the 8 years were not the same, in some months there were higher cases and lower in others regardless of the overall annual cases (see Fig. 5). This observation is supported by the study by Nkurunziza et al. (Nkurunziza et al., 2011) in which the authors found that malaria incidence was positively associated to minimum temperature between two months, and that rainfall and maximum temperature in a given month have a possible negative effect on malaria incidence of the same month. Even if there are some months with high infection rates, same month of the years can have more or less malaria cases if we look at Fig. 5. These different trends in different months show the existence of stochasticity for malaria population in Burundi. This shows how climate change is impacting malaria transmission dynamics in Burundi. But not only climate change can explain those stochastic trends at small time scale, other factors such as socio-economic factors without excluding intensity of prevention interventions such as distribution of ITNs, and medication, can affect malaria transmission dynamics in Burundi.

Table 2
Goodness of fit for malaria cases predicted with generalized linear and artificial neural network models.

Malaria cases	MSRE (GLM)	CV (GLM)	MSRE (ANN)	CV (ANN)
Total	7661	5959	3949	3950
Pregnant women & children (< 5 years)	113	205	486	436
Pregnant women	27	22	24	27
Children (0 and 11 months)	572	662	502	554
Children (12 and 59 months)	535	722	655	748

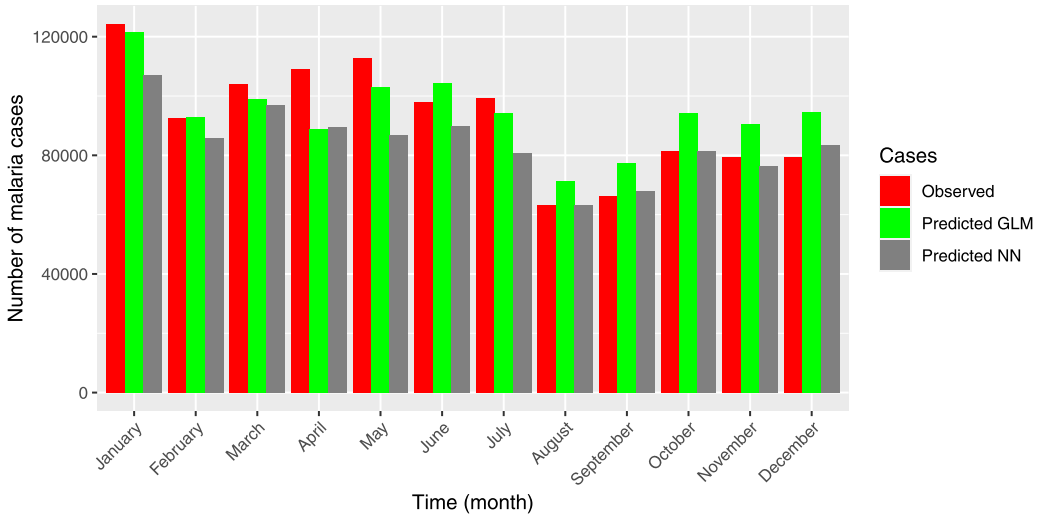


Fig. 7. Observed and predicted monthly total malaria cases in 2017.

One of the limitation in this study is that we could not get granular data at province level to understand truly the effects of socio-economic factors and prevention interventions on malaria cases.

Generalized linear model, and the neural network both estimate number of malaria cases at a certain accuracy. The main difference between these two approaches is that GLM is more transparent than neural network. Despite the fact that they both provide predictors' importance for the model output. The ANN works like a *black-box* which detects multiple nonlinear interactions among a series of input variables before giving the output, whereas the GLM provides details on how the predictors affect the response variable (model output) through their coefficients. We have seen that the neural network model works better for predicting the overall total cases of malaria. However, it does not provide information on the change of transmission dynamics given a change of the magnitude of the effect of predictors (variables) to the outcome (malaria of cases). Thus, for planning purposes, it may be advisable to use neural network to predict overall malaria cases. Generalized linear model gives the effect of each of predictors (variables) and whether it increases or decreases the likelihood of the outcome (malaria cases). Thus, this helps to know where to focus more attention for interventions.

From GLM and ANN results, we have seen that associated parameters to education rates and ITNs had negative estimates, which means they have a decreasing effect on total malaria cases. Therefore, malaria interventions targeting school-children and sensitization in schools to increase awareness, and capacity building for best practices to avoid malaria infection in communities combined to messages targeting school learners and students can also help to reduce malaria infection (Ayi et al., 2010). In a recent systematic review by Cohee et al. (Cohee et al., 2020) for preventive malaria treatment among school-aged children in sub-Saharan Africa, the results showed that such interventions decrease significantly the P. falciparum prevalence, anaemia, and risk of subsequent clinical malaria across transmission settings. Distributing highly effective ITNs to people who know to use it effectively can reduce new malaria infections.

A neural network model is a type of machine learning model that is usually used in supervised learning, whereas a GML model is a type of statistical model. In the literature, not only statistical linear models and machine learning based models are used to estimates malaria cases. Dynamics models (Mandal et al., 2011), agent-based models (Smith et al., 2018), and time series (Hussien et al., 2017; Shi et al., 2020) are also used. However, if we want to understand better the effects of different factors, GLM models and machine learning based models provide more insights. But sometimes, the research questions can be well answered by a given type of model. The new tendency of using different data sources for malaria, which is gaining a lot of attention is mostly driven by agent-based models (Amadi et al., 2021) and machine learning (Lapão et al., 2017). It helps to

understand different factors influencing malaria transmission dynamics. Furthermore, artificial intelligence is also improving diagnostics (Keshavarzi Arshadi et al., 2020; Madhu & Govardhan, 2022).

Malaria is still a public health challenge in Burundi. Deep dive studies can help to ascertain the principal drivers of the transmission which will then present opportunities for tailor made interventions for effective malaria control. Thus, the use of models can help predict malaria epidemics and is proactive for intervention designs. And stratification of populations can help to identify more risky groups, and to tailor interventions. Furthermore, including many variables in the predictive frameworks can increase accuracy and enhance understanding of malaria dynamic and hence tailor adequate intervention measures.

Funding

DN was supported by NRF-TWAS grant number 100014, and KB received support from the EDCTP2 programme supported by the European Union Career Development fellowship TMA2016CDF-1605.

Authors contributions

All authors conceptualized the study. EN, and LM collected and collated the data, DN wrote the code to clean, and analyze the data, DN conducted the analysis and modelling part, KB revised extensively the first draft of the manuscript. All authors wrote and reviewed the manuscript.

Ethical statement

The present work is part of a wide study (Epidemiological and molecular investigation of malaria in Burundi) which was approved by the National Ethics Committee for the protection of human beings subject to biomedical and behavioral research of the Republic of Burundi (CNE/03/2021).

Declaration of competing interest

The authors declare no conflict of interest. The funder had no role in the design of the study; in the analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Acknowledgments

Authors acknowledge the contribution of Emanuel Muema Dominic from South African Centre for Epidemiological Modelling & Analysis (SACEMA) during data analysis and modelling, and Edmund Yamba from Kwame Nkrumah University of Science and Technology for his valuable comments to the early version of the manuscript.

References

Amadi, M., Shcherbacheva, A., & Haario, H. (2021). Agent-based modelling of complex factors impacting malaria prevalence. *Malaria Journal, 20*, 1–15.
Ayi, I., Nonaka, D., Adjovu, J. K., Hanafusa, S., Jimba, M., Bosompem, K. M., Mizoue, T., Takeuchi, T., Boakye, D. A., & Kobayashi, J. (2010). School-based participatory health education for malaria control in Ghana: Engaging children as health messengers. *Malaria Journal, 9*, 1–12.
Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). lme4: Linear mixed-effects models using eigen and s4. In *R package version 1* (pp. 1–23).
Bayoh, M. N., & Lindsay, S. W. (2003). Effect of temperature on the development of the aquatic stages of anopheles gambiae sensu stricto (diptera: Culicidae). *Bulletin of Entomological Research, 93*, 375–381.
Beck, M. W. (2022). *Visualization and analysis tools for neural networks package* (pp. 1–29). R package version.
Beck-Johnson, L. M., Nelson, W. A., Paaijmans, K. P., Read, A. F., Thomas, M. B., & Bjørnstad, O. N. (2013). The effect of temperature on anopheles mosquito population dynamics and the potential for malaria transmission. *PLoS One, 8*, Article e79276.
Bomblies, A., Duchemin, J.-B., & Eltahir, E. A. (2008). Hydrology of malaria: Model development and application to a sahelian village. *Water Resources Research, 44*.
Cohee, L. M., Opondo, C., Clarke, S. E., Halliday, K. E., Cano, J., Shipper, A. G., Barger-Kamate, B., Djimde, A., Diarra, S., Dokras, A., et al. (2020). Preventive malaria treatment among school-aged children in sub-saharan africa: A systematic review and meta-analyses. *Lancet Global Health, 8*, e1499. –e1511.
Density of doctors, nurses and midwives in the 49 priority countries. https://www.who.int/hrh/fig_density.pdf?ua=1, (2010), 2021-05-26.
Dobbs, K. R., & Dent, A. E. (2016). Plasmodium malaria and antimalarial antibodies in the first year of life. *Parasitology, 143*, 129–138.
Doolan, D. L., Dobaño, C., & Baird, J. K. (2009). Acquired immunity to malaria. *Clinical Microbiology Reviews, 22*, 13–36.
Eight facts about Burundi's malaria epidemic. https://www.wvi.org/article/8-facts-about-burundis-malaria-epidemic, (2017), 2021-05-26.
Global technical strategy for malaria 2016–2030. https://www.who.int/malaria/publications/atoz/9789241564991/en/, (2021), 2021-05-26.
Gökhan, A., Güzeller, C. O., & Eser, M. T. (2019). The effect of the normalization method used in different sample sizes on the success of artificial neural network model. *International Journal of Assessment Tools in Education, 6*, 170–192.
Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R Journal, 2*, 30–38.
Gurney, K. (1997). *An introduction to neural networks*. CRC press.
Haque, U., Sunahara, T., Hashizume, M., Shields, T., Yamamoto, T., Haque, R., & Glass, G. E. (2011). Malaria prevalence, risk factors and spatial distribution in a hilly forest area of Bangladesh. *PLoS One, 6*, Article e18908.
Health metrics for Burundi. http://www.healthdata.org/burundi, (2020), 2021-05-26.
Heaton, J. (2008). *Introduction to neural networks with Java*. Heaton Research, Inc.
Hussien, H. H., Eissa, F. H., & Awadalla, K. E. (2017). *Statistical methods for predicting malaria incidences using data from Sudan, Malaria research and treatment 2017*.

Keshavarzi Arshadi, A., Salem, M., Collins, J., Yuan, J. S., & Chakrabarti, D. (2020). Deepmalaria: Artificial intelligence driven discovery of potent antiplasmodials. *Frontiers in Pharmacology, 10*, 1526.
Kipruto, E. K., Ochieng, A. O., Anyona, D. N., Mbalanya, M., Mutua, E. N., Onguru, D., Nyamongo, I. K., & Estambale, B. B. (2017). Effect of climatic variability on malaria trends in baringo county, Kenya. *Malaria Journal, 16*, 1–11.
Lancaster, I. (1980). Relationships between altitude and temperature in Malawi. *South African Geographical Journal, 62*, 89–97.
Lapão, L. V., Maia, M. R., & Gregório, J. (2017). Leveraging artificial intelligence to improve malaria epidemics' response. *Anais do Instituto de Higiene e Medicina Tropical, 16*, 35–39.
Li, T., Yang, Z., & Wang, M. (2013). Temperature, relative humidity and sunshine may be the effective predictors for occurrence of malaria in Guangzhou, Southern China, 2006–2012. *Parasites & Vectors, 6*, 155.
Lok, P., & Dijk, S. (2019). *Malaria outbreak in Burundi reaches epidemic levels with 5.7 million infected this year*. British Medical Journal Publishing Group.
Lozano, R., & Garrido, F. (2015). *Improving health system efficiency (health systems governance and financing)*. World Health Organization. file:///Users/hectorcarrasco/Downloads/WHO_HIS_HGF_CaseStudy_15. 7_eng 20.
Madhu, G., & Govardhan, A. (2022). Artificial intelligence based diagnostic model for the detection of malaria parasites from microscopic blood images. In *Intelligent interactive multimedia systems for e-healthcare applications* (pp. 215–233). Springer.
President's Malaria Initiative. (2016). *Malaria operational plan FY 2017*. United States Agency for International Development (USAID).
Mandal, S., Sarkar, R. R., & Sinha, S. (2011). Mathematical models of malaria-a review. *Malaria Journal, 10*, 1–19.
McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.
Nkurunziza, H., Gebhardt, A., & Pilz, J. (2011). Geo-additive modelling of malaria in Burundi. *Malaria Journal, 10*, 234.
Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling, 178*, 389–397.
Onyango, E. A., Sahin, O., Awiti, A., Chu, C., & Mackey, B. (2016). An integrated risk and vulnerability assessment framework for climate change and malaria transmission in east africa. *Malaria Journal, 15*, 551.
Pascual, M., Ahumada, J. A., Chaves, L. F., Rodo, X., & Bouma, M. (2006). Malaria resurgence in the east african highlands: Temperature trends revisited. *Proceedings of the National Academy of Sciences, 103*, 5829–5834.
Peng, X., Wu, W., Zheng, Y., Sun, J., Hu, T., & Wang, P. (2020). Correlation analysis of land surface temperature and topographic elements in Hangzhou, China. *Scientific Reports, 10*, 1–16.
R Core Team, R. (2018). In *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.
Semakula, H. M., Song, G., Zhang, S., & Achuu, S. P. (2015). Potential of household environmental resources and practices in eliminating residual malaria transmission: A case study of Tanzania, Burundi, Malawi and Liberia. *African Health Sciences, 15*, 819–827.
Shi, B., Lin, S., Tan, Q., Cao, J., Zhou, X., Xia, S., Zhou, X.-N., & Liu, J. (2020). Inference and prediction of malaria transmission dynamics using time series data. *Infectious Diseases of Poverty, 9*, 84–96.
Sinzinkayo, D., Baza, D., Gnanguenon, V., & Koepfli, C. (2021). The lead-up to epidemic transmission: Malaria trends and control interventions in Burundi 2000 to 2019. *Malaria Journal, 20*, 1–7.
Siya, A., Kalule, B. J., Ssentongo, B., Lukwa, A. T., & Egeru, A. (2020). Malaria patterns across altitudinal zones of mount elgon following intensified control and prevention programs in Uganda. *BMC Infectious Diseases, 20*, 1–16.
Smith, N. R., Trauer, J. M., Gambhir, M., Richards, J. S., Maude, R. J., Keith, J. M., & Flegg, J. A. (2018). Agent-based models of malaria transmission: A systematic review. *Malaria Journal, 17*, 1–16.
Stekhoven, D. J. (2015). *missforest: Nonparametric missing value imputation using random forest*. Astrophysics Source Code Library.
Thang, N. D., Erhart, A., Speybroeck, N., Hung, L. X., Hung, C. T., Van Ky, P., Coosemans, M., D'Alessandro, U., et al. (2008). Malaria in central vietnam: Analysis of risk factors by multivariate analysis and classification tree models. *Malaria Journal, 7*, 28.
UNICEF Burundi humanitarian situation report – 31 March 2017. (2017). United Nations For Children (UNICEF)/Burundi.
Wang, X., Yang, B., Huang, J., Chen, H., Gu, X., Bai, Y., & Du, Z. (2016). Iasm: A system for the intelligent active surveillance of malaria. In *Computational and mathematical methods in medicine*.
Weekly Bulletin on outbreaks and other emergencies: Week 27: 01–07 July 2017. (2017). World Health Organization (WHO)/Regional Office for Africa.
World malaria report 2017. (2018). Geneva: World Health Organization (WHO).
World health organization's global health workforce statistics. https://data.worldbank.org/indicator/SH.MED.PHYS.ZS, (2017), 2021-05-26.