IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

Predicting Malaria Incident Using Hybrid SARIMA-LSTM Model

J. S. Adeyeye[1*], E. B. Nkemnole[2]

1,2 Department of Statistics, University of Lagos, Nigeria.
* Corresponding author: jadeyeyesunday@yahoo.com, enkemnole@unilag.edu.ng

Article Info
Received: 06 January 2023 Revised: 05 August 2023
Accepted: 07 August 2023 Available online: 20 August 2023

Abstract

Malaria remains a significant global health concern, particularly in regions with high transmission rates. Accurate and timely prediction of malaria incidence can assist health authorities and policymakers in implementing effective prevention and control measures. However, because data are in limited supply, most of the relevant research studies concentrated on monthly or quarterly data. This study proposes a hybrid forecasting model combining Seasonal Autoregressive Integrated Moving Average (SARIMA) and Long Short-Term Memory (LSTM) neural networks to predict malaria incidence. The hybrid approach enhances accuracy and robustness by capturing historical data's temporal dependencies and seasonal patterns. The methodology involves collecting historical malaria incidence data, preprocessing it, fitting SARIMA models, extracting residuals, and training LSTM neural networks on residuals. These models capture nonlinear and complex data components, making accurate predictions and capturing long-term dependencies. After training, the hybrid SARIMA-LSTM model is created by combining the predictions from both models. This integration ensures that both the temporal and nonlinear patterns are considered, leading to improved forecast accuracy. Finally, the model is evaluated using appropriate performance metrics, such as mean absolute percentage error (MAPE) or root mean square error (RMSE). The hybrid SARIMA-LSTM model outperforms SARIMA and LSTM in predicting malaria incidence and its accuracy was evaluated through comparisons with other forecasting methods. It captures temporal and nonlinear patterns, enabling timely resource allocation, intervention planning, and proactive measures for improved control and prevention efforts.

Keywords: Malaria, Long Short-Term Memory, Seasonal Autoregressive Integrated Moving Average, Seasonal Autoregressive Integrated Moving Average-Long Short-Term Memory, Predictive Accuracy.
MSC2010: 26A18.

1 INTRODUCTION

Malaria is a life-threatening infectious disease caused by parasites transmitted to humans through the bites of the infected mosquito vector. The world malaria report released in December 2021

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

reflects the global malaria community's unique challenges. The report showed the devastating toll of malaria, with an estimated 627,000 losing their lives to the disease in 2020 [1]. Sub-Saharan African countries are believed to be at the epicenter of this malady with Nigeria, a West African country, leading the trajectory. According to World Health Organization 's report on Malaria, Nigeria alone accounted for 31% and 31.3% of global malaria deaths in 2021 and 2022 respectively. The most vulnerable to this fatality are children under the age of five. Pregnant women are also not spared as it causes a high rate of miscarriages. It is a major concern for all and why an age-long disease that is both preventable and treatable could pose a major public health challenge to Africa 's most populous nation. This work aims to use two-time series statistical approaches to look into and predict the prevalence of malaria in Nigeria, as well as to create a hybrid model and evaluate it against more conventional models. In the literature, many authors used single traditional time series to predict incidences of malaria but it was discovered that a single model cannot effectively capture all the properties of the data structure unless the use of stacking architecture which involves the combination of distinct algorithms and models is used, Wang et al. [2]. LSTM has the advantage of being able to retain information for a very long period, unlike recurrent neural networks (RNN), and has a wide range of parameters such as learning rates and 'input and output biases ', whereas classical time series have no further opportunity for fine changes. ARIMA, on the other hand, predicts future values based on past values by smoothing time series data with lagged moving averages. It performs well on short-term forecasts but poorly on long-term projections. When forecasting malaria incidence, combining LSTM and ARIMA models would surely produce a superior outcome. Several statistical methods have been used to predict malaria incidence in previous studies. Wangdi et al. [3] developed a temporal model for forecasting and predicting of malaria infections using time-series and Arimax analyses in the endemic districts of Bhutan. The study revealed that the ARIMA model performed better than the ARIMAX model and can therefore be employed for planning and managing malaria prevention and control programs in Bhutan. Okoli, [4] carried out an investigation on the incidence and mortality rate of reported cases of malaria in Anambra state, Nigeria. The study used two-way ANOVA, multiple comparison tests, the test of equality of proportions, runs test, and trend analysis to carry out the study and found out that: the mean incidence of malaria in Anambra state differs significantly across age groups, the mean mortality is equal across the years, the mean mortality differs across age groups, incidence, and mortality was found to be equal between males and females, among others. Egbuche, et al. [5] determined the composition of species of anopheles and some climatic factors that influence their survival and population abundance in Anambra East LGA, Nigeria. Four Anopheles species: An. Gambiae s, An. funestus group, An. moucheti, and An. nili were identified in the study of 8181 female anopheles mosquitos comprising 4127 larvae and 4054 adults. Kassa et al. [6] assessed the control measures and trends of malaria in Burie-Zuria district, Ethiopia by undertaking descriptive cross-sectional control measures and found that the attack rate was higher among children that are less than 5 years old when compared to other age groups with no sex difference. Olawale and Donaldson [7] worked 'on time domain analysis of malaria morbidity in Nigeria '. ARIMA model was built for analyzing the secondary data collected on the incidence of malaria. It was observed that there is going to be a steady increase in malaria prevalence. Santosh et al. [8] proposed a novel scalable framework to predict the instances of malaria in selected geographical locations. The study employed satellite data and clinical data along with a long short-term memory (LSTM) classifier to predict malaria abundances in the state of Telangana, India. It was revealed that the Apache Spark-based LSTM presents an effective strategy to identify locations of endemic malaria. Permanasari et al. [9] analyzed and presented the use of the seasonal autoregressive integrated moving average (SARIMA) method for developing a forecasting model that can support and provide a prediction of the number of malaria incidences in humans. Tuan Tran et al. [10] predicted the P. falciparum gene transcription during its blood stage life cycle, implementing a well-tuned recurrent neural network with gated recurrent neural units. The results of the study showed a high level of accuracy in being able to predict and forecast the expression levels of the different genes. Olatayo and Adedotun [11] showed the efficiency of the different methods used to test and estimate fractional parameters in the fractionally integrated autoregressive moving average (AFRIMA) model. Faniran and Ayoola [12]

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

designed a study to model the transmission dynamics of malaria, taking into consideration some infectious humans who do not comply with drug. Iyare and Akhaze [13] proposed an ordinary differential equation co-infection model of tuberculosis-lymphatic filariasis with 17 mutually disjoint compartments.

2 Materials and Methods

2.1 The SARIMA model

According to Makridakis and Hibon [14], the ARMA model has evolved into the ARIMA model. The moving average model MA(q) and the autoregressive model AR(p) are combined to form the ARMA model. The AR and MA models are predicted in a particular order using the ARMA model. Only data with normality are suitable for the application. ARMA is used for modeling when a time-series graph lacks a consistent pattern and the Auto-Correlation Function (ACF) steadily declines. Time series data containing anomalies are transformed or differentially normalized before being modeled by ARMA. It is known as the ARIMA model. The average regression model, or AR(p) model, is being used here with the assumption that the historical data from Y_t can adequately describe the current time series data, or $Y_(t-p)$ to $Y_(t-1)$. It is predicated on the notion that the data from the most recent time series are reliant on the data from the most recent series. The current data is independent of the past and turns into white noise time series data if there is even a slight dependency between them. The more dependent on the past the current data is, up to and including a random walk, the stronger the dependency on the past. By examining the autocorrelation with the past, the AR(p) model examines the properties of the target time series data. It implies that the data from time t influence the data now. ACF rapidly declines in the graphs of the Autocorrelation Function (ACF) and Partial Auto-Correlation Function (PACF), while PACF has a cut point at a specific point. The term ACF refers to the correlation between data that are spaced by k periods. In other words, the ACF shows the correlation's order based on the time difference. The Partial Auto-Correlation Function (PACF), in contrast to the ACF, is a pure correlation coefficient between two variables that is determined after removing the influence of all potential differences between the observed values. The PACF transforms into an AR(1) model if k = 2 is the cut-off point.
A general regression model of AR(p) is as follows:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \varepsilon_t \tag{2.1}$$

The equation (2) below is an AR(p) model expressed in ARMA (Autoregressive Moving Average) form. In this form, μ represents the mean of the time series, and $(Y_{t-1}-\mu), (Y_{t-2}-\mu), ..., (Y_{t-p}-\mu)$ are the deviations from the mean at each lagged time point. The coefficients $\alpha_1, \alpha_2, ..., \alpha_p$ represent the autoregressive coefficients, p is the autoregressive order and ε_t is the white noise with mean 0 and variance σ^2.

By centering the lagged values around the mean μ, the ARMA form can simplify the interpretation of the model. It allows us to explicitly see the effect of each lagged value on the current value, relative to the mean of the time series.

$$Y_t = \mu + \alpha_1(Y_{t-1} - \mu) + \alpha_2(Y_{t-2} - \mu) + ... + \alpha_p(Y_{t-p} - \mu) + \varepsilon_t \tag{2.2}$$

The current time series data is made up of a weighted average of historical residuals, and the MA model is a moving average procedure. The current data is defined as the mean value of previous white noise because the residual term is white noise. The MA model based on the sum of them has an average regression characteristic because the white noise has a high normality and high average regression characteristic. ACF has a breaking point, while PACF exhibits a sharp decline.
The moving average model MA(q) is a model of weighted linear combination with white noise t, in

IJMSO
International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

contrast to the autoregressive model AR(p). The current time series data Y_t can be expressed by continuous error terms $\varepsilon_{t+1}, \varepsilon_{t+2}, \varepsilon_{t+3}, ..., \varepsilon_{t+q}$. The general form of MA(q) model is as follows:

$$Y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - ... - \theta_q\varepsilon_{t-q} \tag{2.3}$$

Where, ϵ_t represents the white noise with mean 0 and variance σ^2, θ_q the moving average coefficient, q the order of moving average. Hence, MA(1) is expressed as $Y_t = \varepsilon_t - \theta_1\varepsilon_{t+1}$.
It is challenging to estimate general time series data using only AR(p) or MA(q). Then, Autoregressive Moving Average (ARMA), which combines the best features of both models, is applied. The ARMA model combines the AR and MA models and assumes that the function of historical time series data and historical residuals determines the current time series data. The ARMA model has the average regression characteristic, just like the AR and MA models do.
The average regression characteristic of the AR, MA, and ARMA models makes them appropriate for time series analysis, which always has normality for all parameter values. In comparison to existing AR or MA models, the ARMA model approximates the value relatively more accurately and quickly with fewer parameters.
The values of ARMA(1,0) and ARMA(0,1) are equal to AR(1) and MA(1), respectively, because ARMA is a mixed model of the AR and MA models. The ARMA model's general formula is as follows:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - ... - \theta_q\varepsilon_{t-q} \tag{2.4}$$

The majority of time-series data typically lacks normality and exhibits rising trends or rising variance over time. Because of the unstable time series, the predicted value is no longer valid because the mean and variance of the time series change over the course of time. AR, MA, and ARMA models cannot be used to analyze such time series data. As a result, the data needs to be normalized before being converted to a time series.
In order to convert in accordance with the properties of the data, log transformation, difference, and seasonal differences are carried out. The ARIMA model is used to analyze the time series after it has been normalized.
To analyze time series models with seasonal patterns, one can use the regression model using indicator functions and trigonometric functions or Winters' seasonal exponential smoothing; however, these techniques can only be applied when the seasonal time series data are independent of one another. However, since time series data are typically correlated with one another, the ARIMA model is the most appropriate.
Even if the data itself lacks normality or average regression characteristics, some data may have a time series average regression characteristic after difference. The differential time series are used in the ARIMA model, which is an ARMA model. The ARMA model and the ARIMA model with difference value 0 are equivalent.
The process of differencing is as follows:

$$\nabla Y_t = (1 - B)Y_t = Y_t - Y_{t-1} \tag{2.5}$$

$$\nabla^2 Y_t = (1 - B)^2 Y_t = (1 - 2B + B^2)Y_t - 2Y_{t-1} + Y_{t-2} \tag{2.6}$$

Where, B is the backshift operator which means $B^j Y_t = Y_{t-j}$
The process of deducting the prior data from the initial data until the time series data are normal is what makes the difference. Three orders of ARIMA exist: p, d, and q. These orders are expressed as ARIMA(p,d,q), where p is the number of autoregressive terms, d is the number of nonseasonal differences required for stationarity, and q is the number of lagged forecast errors in the prediction equation.
When the time series data show seasonal trends, seasonal ARIMA is generally used. The seasonal, autoregressive, intergraded, and moving average components are combined in the SARIMA. The model assumes that malarial incidence data include trends, seasonal components, and irregular terms. The following steps are taken in this study to develop the SARIMA models. To begin with, the Osborn, Chui, Smith, and Birchenhall (OCSB) test is employed to establish the seasonal

IJMSO
International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

differencing order. The KPSS unit-root test is then used to determine the order of differencing. The model space is then traversed using stepwise processes to identify the order of autoregressive and moving average terms, p, q, P, and Q. Finally, the optimal model is chosen using goodness-of-fit tests based on AIC and estimated residuals (Taiwo et al. [15]; Onasanya et al. [16]). The general formula of SARIMA(p,d,q) (P,D,Q)s is as follows:

$$SARIMA = (p, d, q) \times (P, D, Q)_s \tag{2.7}$$

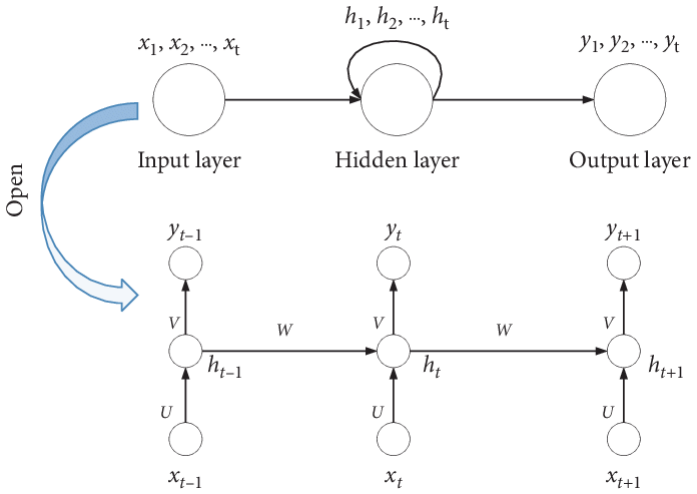
Where the additional features are $(P, D, Q)_s$ and they are described as follows:
P= Order of seasonality, mathematically, the SARIMA model is written as:

$$\Phi(B^s)\phi(B)\triangle_s^D\triangle^d X_t = \Theta(B^s)\theta(B)\varepsilon_t \tag{2.8}$$

Where $\varepsilon_t = whitenoise$, B is called the back shift operator, $\Phi(B^s)$ is the seasonal Autoregressive (AR) coefficient, $\Phi(B)$= the seasonal Autoregressive (AR) coefficient, $\triangle^D$= the number of nonseasonal difference d, $\triangle_s^D$= the number of seasonal difference D, X_t=the observation of seasonal time series without normality, $\Theta(B^s)$= seasonal moving average coefficient (MA), $\theta(B)$= nonseasonal moving average coefficient (MA) and ε_t= the error term or white noise. If the order of seasonal time series model is zero, it is the same with ARIMA.

2.2 LSTM (Long short-term memory) model

This is another statistical method introduced to predict the incidence of malaria as this method is more robust than ARIMA model in so many ways. The approach in determining the future trends of this malaria incidence is more realistic than ARIMA model and its prediction gives accurate result. The word LSTM means long short-term memory, which is a network from a recurrent neural network. One of the main reasons of introducing LSTM is because the malaria incidence is a time series, which occurs in periodical manner, and the nature of this series is a number. Now we want to predict the malaria incidence where the gradient obtained is controlled or overcome and most important to capture the linear long-term dependencies in the sequence of malaria incidence where each neuron accommodates a memory room, which has the capacity of storing previous information used by the recurrent neural network or forgetting if there is necessity. In inclusion to the memory cell, the long short- term memory cell contains what we call an input gate, output gate and forget gate where each gate in the memory cell sustains the current input x_t, the hidden state $h_(t-1)$ at the earlier instant and the state information symbolized as $c_(t-1)$ of the internal cell memory to execute different action and also to dictate whether to activate using logic task. The output at time t, the state h_t of the unit and the hidden state at the time t_1 are all set on non-linear activation of tanh and the available information of output gate. The following figures represent the structures of recurrent neural network and that of a long short-term Memory



IJMSO
International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

Figure 1: Structure of a recurrent neural network

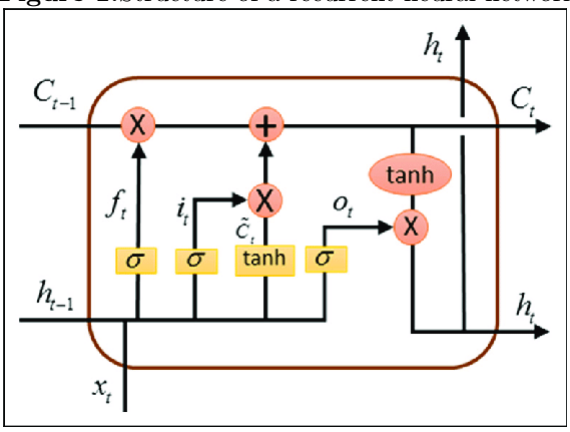


Figure 2: Structure of a long short-term Memory

Step1: Some informations that are not needed in the LSTM are identified first and once they are identified, they are removed or thrown away from the LSTM cell through the sigmoid gate layer also known as the forget layer, it is defined mathematically as:

$$f_t = \sigma[w_f(h_{t-1}, x_t) + b_f] \tag{2.9}$$

$f_t = \sigma(w_{fx}x_t + w_{fh}h_{t-1} + b_f)$
Where w_f= the weight assigned, x_t= the input, h_{t-1}= the output from old time stamp and b_f= the bias.

$$i_t = \sigma[w_i(h_{t-1,x_t}) + b_i] \tag{2.10}$$

$i_t = \sigma(w_{ix}x_t + w_{ih}h_{t-1} + b_i)$

$$\hat{c}_i = tanh[w_c(h_{t-1}, x_t) + b_i] \tag{2.11}$$

The function of the new cell state is described below:

$$c_t = f_t * c_{t-1} + i_t * \hat{c}_i \tag{2.12}$$

$g_t = \phi(w_{gx}x_t + w_{gh}h_{t-1} + b_g)$
The output gate layer will determine which fragment of the cell state will be the output. The output function is expressed mathematically as:

$$o_t = \sigma[w_o(h_{t-1}, x_t) + b_i] \tag{2.13}$$

$o_t = \sigma(w_{ox}x_t + w_{oh}h_{t-1} + b_o)$
$s_t = g_t.i_t + s_{t-1}.f_t$
$h_t = \phi(s_t).o_t$

$$h_t = o_t * tanh(c_t) \tag{2.14}$$

Where $w_{fx}, w_{fh}, w_{ix}, w_{ih}, w_{gh}, w_{ox}$ and w_{oh} are weight parameters for the corresponding output of the network activation function; σ and ϕ are sigmoid functions and tanh(.), respectively. The sigmoid function with an output range of [0,1] works as a soft switch for the forget gate (f_t), input gate (i_t), input node (g_t) and output gate (o_t)
This means that it is a decision-making point to determining whether the signal/sequencing data

IJMSO
International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

should pass the gate or not.
Thus, all gates (forget gate, input gate, input node and output gates), are directly depended on the current x_t and previous output H_{t-1}

2.3 SARIMA-LSTM

There are linear and nonlinear relationships between time series data. Although statistical methods are effective at managing linear relationships in time series, they are incapable of handling nonlinear relationships (Martnez et al. [17]; Adesina et al. [18]). Neural network approaches, on the other hand, can model both linear and nonlinear relationships, but they require careful parameter selection and a lengthy training period. However, a hybrid model is implemented to compensate for these deficiencies and enhance performance. Seasonal time series predictions made by the SARIMA model have shown to be accurate. The benefits of the SARIMA and LSTM models are combined in the suggested SARIMA-LSTM model. SARIMA is employed to record the seasonal and trend components of malaria incidence. The input layer of the LSTM model receives residuals produced by the SARIMA model. The SARIMA-LSTM method takes advantage of the SARIMA model's ability to forecast outcomes and the nonlinear model's capacity to further minimize residuals. For the SARIMA model and LSTM model in this work, each batch of collected time series data is divided into two data sets: the training data set (in-sample data) and the testing data set (out-of-sample data). The training data utilized in the SARIMA-LSTM model study are evenly split into two parts: the first 50% of the training data are used to determine the size of the SARIMA model's rolling windows, which is then used to make one step forward rolling predictions and obtain matching error. To construct predictions and obtain related residual errors, the estimates of the SARIMA model are combined with the remaining testing data. These mistakes represent the accuracy of predicting models. The input data sets of the LSTM model are residual errors. The first 54.5% of the residual error data sets are utilized as the training data for the LSTM model, and these first 54.5% of the residual error data sets correspond to the middle 35.3% of the original data. Residual error data sets are likewise separated into training data and testing data. The LSTM model tests its performance using the remaining 45.5% of the residual error data sets, which correspond to the final 29.4% of the original data. The search range of the initial hidden layers and initial number of hidden neurons are chosen at 1 to 3 and 1 to 30, respectively, because a better prediction result was discovered between the hidden layers of 1 to 3 and hidden neurons of 1 to 30. First, the quantity of hidden layers is fixed, and the number of hidden neurons is searched within a range of 1 to 30. The value of 30 training RMSE (root mean square error) is calculated, and the number of hidden neurons increases by one for each cycle. The number of hidden layers then rises by one, and the range of hidden neuron searches is set at 1 to 30. Up till 3 * 30 training RMSE values are calculated, the values of training RMSE are calculated as the number of hidden layers and neurons changes.

2.4 Evaluation Metrics

Despite the fact that there have been numerous suggested metrics for evaluating time series models, we choose to concentrate on the Mean Average Percent Error, Root Mean Squared Error, and Mean Average Error. $\hat{y}^t$ represents the model's forecast for each of the following measures of error, while y^t represents the actual value at time t.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{y}^t - y^t|}{y^t} \tag{2.15}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}^t - y^t)^2} \tag{2.16}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}^t - y^t| \tag{2.17}$$

IJMSO
International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

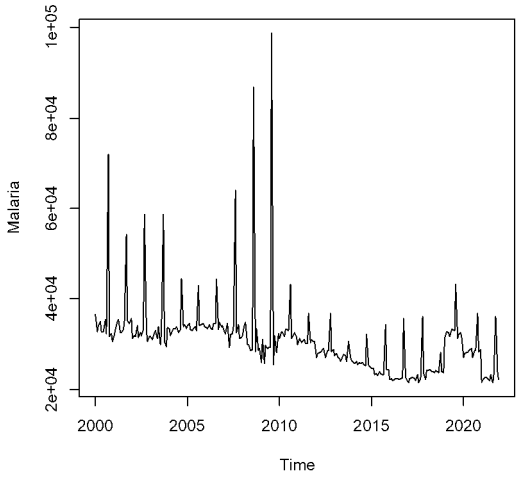
The percentage difference between the prediction and the actual can be calculated using the MAPE measurement, which is frequently used in literature. However, it seems that MAPE performs poorly on data sets with a lot of low or zero-valued scores of 0, as MAPE approaches infinity as the actual value approaches 0. As a result, in addition to reporting MAPE scores, we also took into account the MAE as a more broadly applicable estimator of the typical difference between the prediction and the actual value. The worst possible MAE on a data set with values ranging from 0 to 100 is 100, while a naive model with a constant value of 50 would achieve an MAE of 50 or less. Each model was evaluated, and the corresponding RMSE distribution was utilized to produce 95% confidence intervals (CI) and p-values comparing significant differences to determine the statistical significance of the reduction in RMSE in models compared to LSTM.

3 Results

3.1 Data Source

The data used for the research study is a monthly case of malaria from January 2000 to December 2021, and these data was obtained or coined from World Bank data bank (https://www.datacatalog.worldbank.org). There are some specific unreported cases of malaria across Nigeria. These data obtained are available from cases reported nationwide across all the states in Nigeria over the sample periods of 21 years. The training data points used for building the ARIMA model started from January 2000 to December 2021 and a short-term forecast was implemented to determining whether the accuracy of the ARIMA Model built.

In this section, the performance of the proposed hybrid model is evaluated using the latest available public data on malaria. The R package version 3.6.3. was used for this study. Plotting time series data is the first and most crucial step in creating a time series model. The main goal is to examine any elements that might show up in the time series. The time plot of malaria incidence is shown in Fig 4. Malaria incidence exhibits seasonal components and random components because the seasonal fluctuations and random fluctuations are partially or roughly constant in size. Fig 3 shows a downward trend from the first quarter of 2010 to December 2018, and this can be seen from Fig 4 which displays the decomposition graph of malaria incidence. The peak and trough of malaria occurred mostly in January having a seasonal index of value (-2464.5) and while high malaria incidence occurred in August. The series is hereby described using an additive seasonal time series model. It is this month that the mosquitoes breed more, and the spread of malaria get increased and then decrease after during the dry season.



IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

Figure 3: Time plot of Malaria

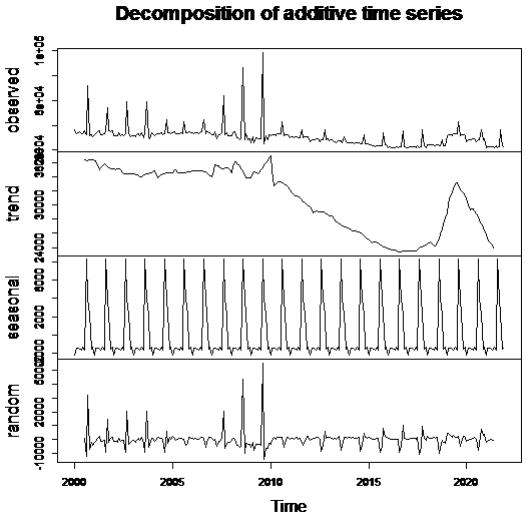


Figure 4: Decomposition of Malaria Data

Since malaria incidence contains seasonal factors, hence SARIMA model was built based on the following series of orders with a period of seasonality. Table 1 shows the different types of SARIMA models built. The best model was the SARIMA model (1, 2, 2) because it has the smallest Alkaike information criteria of value **3318.27**

The SARIMA model-(1, 2, 2) attained this smallest AIC after taken the differencing order to be

	(1,0,1)	**(1,1,1)**	**(0,1,1)**	**(1,1,0)**	**(1,1,2)**	**(2,1,1)**	**(2,2,2)**	**(1,2,2)**	**(2,2,1)**
AIC	3354.06	3334.65	3349.69	3341.86	3329.38	3325.91	3364.22	**3318.27**	3342.68
Nobs	267	266	266	266	266	266	265	265	265
Loglik	-2754. 48	-2747.63	-2783.57	-2699.10	-2763.44	-2784.90	-2774.62	-2776.51	-2792.77

Table 1: Varieties of SARIMA model

two so as to attain stationary, the maximum number of p order for autoregressive is 1 and maximum order of q for moving average is 2. Since the model is determined, the following table 2 shows the parameter values of SARIMA model-(1, 2, 2)

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

	AR1	MA1	MA2	SAR1	SMA1
Coeff.value	-0.849	-1.67	0.826	0.755	0.1971
S.E	0.0750	-	-	0.0651	0.0715

Table 2: Parameter values of SARIMA model-(1,2,2)

Table 2 shows that the values of every coefficient are all less than 1, indicating that the stationary and invertibility conditions were all met. The residuals of the top model found, SARIMA, were used to support the model's estimated predictive ability using the correlogram tool.

X^{2VALUE}	DF	P_{VALUE}
5.2609	10	0.8731

Table 3: Ljung Box of Residual obtained from SARIMA model-(**1,2,2**)

We conclude that the residuals of the SARIMA model - are not serially correlated and that the SARIMA model - is a good model to predict or forecast the incidence of malaria because, according to table 3, the $P_{value} = 0.8867$ is higher than the exact level of significance-alpha value 0.05, indicating that the residuals are independent to each other, which simply means that they are not related. The forecast for malaria incidence in Fig. 5 for the following five years shows that the incidence of malaria will increase from January 2019 to December 2023. The prediction intervals for malaria incidence were presented or reported at 80%, 85%, 90%, and 95%.

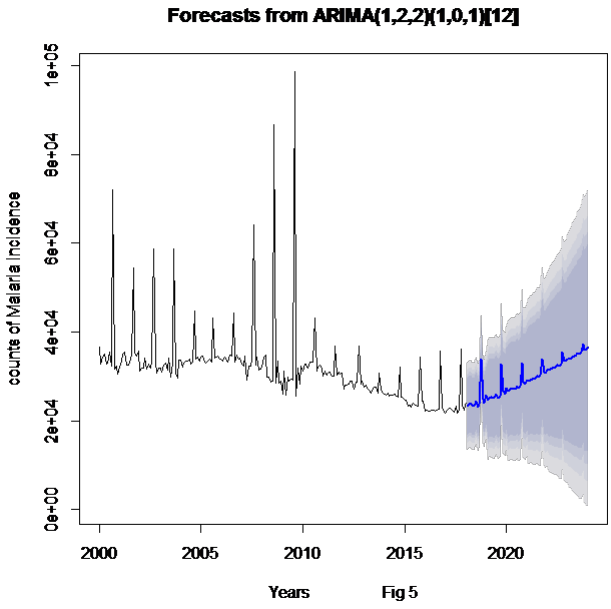


Figure 5: Forecast from SARIMA (1, 2, 2),(1,0,1),(1,2)

Different LSTM models were created using various activation methods, optimizers, and learning rates. Tables 4 and 5 show the loss value and accuracy value of the LSTM models under the aforementioned conditions:

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

			Activation type			
			Softmax	**Relu**	**LR**	**RMSE**
OPTIMIZER	**ADAM**	**Loss value**	2.55	-	0.01	20.653
		Accuracy value	0.3884	-		
	SGD	**Loss value**	0.094	1258.68	0.01	6.3498
		Accuracy value	0.9354	0.1259		

Table 4: The Loss and Accuracy value of malaria LSTM model under learning rate 0.01

			Activation type			
			Softmax	**Relu**	**LR**	**RMSE**
OPTIMIZER	**ADAM**	**Loss value**	143.9	-	0.02	14.36
		Accuracy value	0.5379	-		
	SGD	**Loss value**	0.004	1173.2	0.02	**6.2143**
		Accuracy value	0.9278	0.3835		

Table 5: The Loss and Accuracy value of malaria LSTM model under learning rate 0.02

According to tables 4 and 5, the accuracy value of the malaria LSTM model under the optimizer "SGD" remained constant with respect to the training rate and learning rate for the malaria data, and both models produced the lowest RMSE, with its counterpart "ADAM" optimizer varying with activation types.

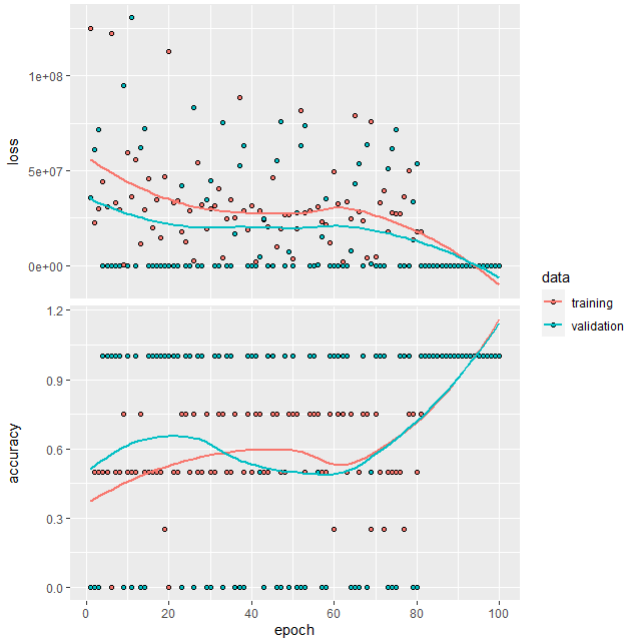


Figure 6: The training and validation data described under the epoch of 100 and learning rate of 0.02

In figure 6, the epoch parameter was set to 100 for each model's training, and dropout was used to reduce overfitting and enhance the model's performance. Srivastava et al.

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

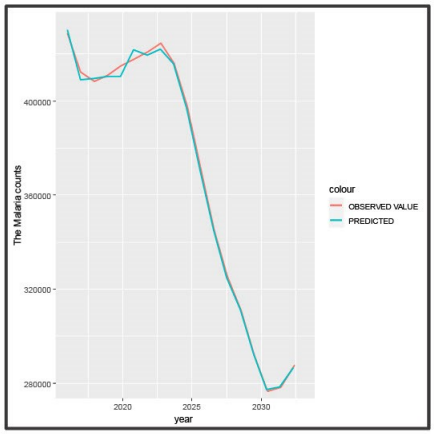


Figure 7: The plot of observed value of malaria Incidence against predicted value from SARIMA-LSTM

Different Approaches	RMSE	MAPE
ARIMA	13784.63	0.4551
SARIMA	11637.31	0.4441
Naive-1	18901.53	0,6658
Mean	16266.06	0.5749
Seasonal Naive	23896.55	0.6961
SES	18901.59	0.6658
Holt	47442.27	1.1231
LSTM	16742.28	0.6286
SARIMA-LSTM	11254.24	0.4228

Table 6: RMSE and MAPE for Forecasting Malaria Incidence Using Different Approaches

Table 6 compares the model prediction comparison results. In both forecasting scenarios, the RMSE and MAPE of LSTM-SARIMA are less than those of SARIMA and LSTM, indicating that LSTM-SARIMA outperforms other models.

4 Discussion

The SARIMA (1, 2, 2) model was created, and it was the best model for capturing the structure or pattern of malaria incidence in the areas where it was diagnosed. One layer with a lot of neurons was used to build the LSTM model, which was optimized in relation to learning rate (for

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

training the data) for model comparisons. The structure and pattern of malaria incidence were produced by activation - "softmax" and optimizer - "sgd" with learning rate (0.02). The LSTM model's predictive accuracy was 92.3%. The training set was used to train the LSTM model and execute a one-step rolling forecast until the final values are predicted. The final step is to calculate the predicted value of the SARIMA-LSTM model by combining the predicted values of the LSTM model and the SARIMA model for the period from January 1, 2022 to December 30, 2030. The results of the RMSE and MAPE for forecasting Malaria occurrence in the country are summarized in Table 6. The Nave-1 and SARIMA models have higher RMSE and MAPE than the proposed SARIMA - LSTM model. It demonstrates the veracity of the SARIMA-LSTM forecasting model. Table 6 demonstrates that the SARIMA-LSTM model has the lowest RMSE and MAPE. RMSE and MAPE for SARIMA-LSTM models are 11254.24 and 42.28%, respectively. These outcomes are the lowest compared to other benchmarking models. This study demonstrates the accuracy of the applied SARIMA-LSTM model, as it is able to reduce error levels. Based on the comparisons and results presented, the SARIMA-LSTM models can improve the error rate and accuracy rate; therefore, the hybrid model is proposed and recommended in this study.

5.0 Conclusion

For malaria control and intervention, early diagnosis is crucial. This can lessen the disease's devastating effects on morbidity and mortality. The ability to predict future trends in disease incidence will greatly improve national control and prevention strategies. With little input, the model offers a way to better understand the dynamics of malaria in an environment with limited resources, producing a forecast that can be applied to sub-district-level public health planning. In addition, MDG 6 attempts to fight diseases like malaria, AIDS, and other illnesses. Malaria and other illnesses affect food and nutrition security, agricultural production, and rural development directly and indirectly. Malnutrition and food hardship can also make a person more susceptible to illness. Incorporating HIV, malaria, and other diseases into food, nutrition, and agricultural policies and programs require the support of FAO policymakers and program planners. The major goal of the study was to use two-time series statistical approaches to look into and predict the prevalence of malaria in Nigeria, as well as to create a hybrid model and evaluate it against more conventional models. The information utilized was obtained from the World Bank data bank (https://www.datacatalog.org) between January 2003 and December 2019. We conclude that malaria tends to be low during the dry season (periods in January) and high during the rainy season (periods in August) due to the presence of seasonal variability. The findings show that for the years under study, the SARIMA-LSTM model's forecasting accuracy outperform that generated by the SARIMA and LSTM models separately. The suggested framework is based on conventional time series that include random, seasonal, and trend elements. While LSTM can extract significant patterns from the random component, SARIMA is utilized to address the seasonal and trend components. In comparing the output of the SARIMA model with that of the SARIMA-LSTM model in table 7 reveals that the SARIMA-LSTM model outperforms the SARIMA model in terms of prediction performance. Due to the presumption that the data set contains both linear and nonlinear characteristics, the SARIMA model is unable to capture all of the data features in the data set. As a result, some nonlinear data features are present in the residuals of the SARIMA model that cannot be captured by the SARIMA model; as a result, the nonlinear feature is then captured by the LSTM model to make a further prediction. The findings of the study confirm our hypothesis. Based on the findings of this study, it is advised that the developed model be taken seriously by the government, health-related NGO's, and policy makers to enable them to administer adequate and prompt malaria control and preventive measures in Nigeria.

Acknowledgments

The authors wish to thank the anonymous reviewers of this paper and state categorically that their careful reading and insightful comments have greatly contributed to its presentation.

Declarations of interest: none.

IJMSO

International Journal of Mathematical Sciences and Optimization: Theory and Applications
Vol. 9, No. 1, pp. 123 - 137
https://doi.org/10.5281/zenodo.8269739

References

[1] April Monroe, Nana Aba Williams, Sheila Ogoma, Corine Karema and Fredros Okumu. (2022): Reflections on the 2021 world malaria report and the future of malaria control. Malaria Journal. 21: 154

[2] Mengyang Wang, Hui Wang, Jiao Wang, Hongwei Liu, Rui Lu, Tongqing Duan, Xiaowen Gong, Siyuan Feng, Yuanyuan Liu, Zhuang Cui, Changping Li and Jun Ma. (2019): A novel model for malaria prediction based on ensemble algorithms. PLOS ONE | https://doi.org/10.1371/journal.pone.0226910

[3] Kinley Wangdi, Pratap Singhasivanon, Tassanee Silawan, Saranath Lawpoolsri, Nicholas J. White and Jaranit Kaewkungwal.(2010): Development of temporal modelling for forecasting and prediction of malaria infections using time-series and Arimax analyses: A case study in endemic districts of Bhutan. Malaria Journal, 9: 251.

[4] Okoli, C. N.(2019): An investigation on incidence and mortality rate of reported cases of malaria in Anambra state, Nigeria. COOU Journal of Physical Sciences. 1(2).

[5] Egbuche, C. M., Onyido, A.E., Umeanaeto, P. U., Nwankwo, E. N., Omah, I. F., Ukonze, C. B., Okeke, J. J., Ezihe, C. K., Irikannu, K. C., Aniekwe, M. I., Ogbodo, J. C and Enyinnaya, J. O.(2020): Anopheles species composition and some climatic factors that influence their survival and population abundance in Anambra East LGA, Anambra State, Nigeria. Journal of parasitology. ISSN 11174145, Volume 41[2].

[6] Addisu Workineh Kassa, Mulugojjam Tamiru and Addisu Gize Yeshanew.(2015): Assessment of control measures and trends of Malaria in Burie-Zuria district, West Gojjam zone, Amhara region, North West Ethiopia. Hindawi Publishing Corporation. Volume 2015, Article ID 302194, 5 pages, http://dx.doi.org/10.1155/2015/302194

[7] Adeboye Nureni Olawale and Ezekiel Imekela Donaldson. (2018): On time domain analysis of malaria morbidity in Nigeria. American Journal of Applied Mathematics and Statistics. Vol.6, No. 4, 170-175: DOI: 10.12691/ajams-6-4-7

[8] Thakur Santosh, Dharavath Ramesh and Damodah Reddy. (2020): LSTM based prediction of malaria abundances using big data. Computers in Biology and Medicine. 124(3):103859, DOI: 10.1016/j.compbiomed.2020.103859

[9] Adhistya Ema Permanasari, Indriana Hidayah and Isna Alfi Bustoni. (2013): SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of malaria incidence. ICITEE 2013 Yogyakarta, 7-8, ISSN: 2088-6578

[10] Tuan Tran, Banafsheh Rekabdah and Chinwe Ekenna. (2021): Deep learning methods in predicting gene expression levels for the malaria parasite. Frontiers in genetics, 22, doi: 10.3389/fgene.2021.721068

[11] Olatayo T. O. and Adedotun A. F. (2014): On the test and Estimate of Fractional Parameter in AFRIMA Model. Applied Mathematical Science, Vol. 8, No.96, 4783-4796.

[12] T.S Faniran and Ayoola E.O. (2019): Mathematical Analysis and Basic Reproduction Number for the Spread and Control of Malaria Model with non-Drug Compliant Humans. International Journal of Mathematical Analysis and Optimization: Theory and Applications. Vol.2019, No. 2, pp. 558-570

[13] Egberanmwen Barry Iyare and Rosemary Uwaila Akhaze. (2022): Derivation of the Reproduction Number and Simulation of the Tuberculosis-Lymphatic filariasis Co-infection Model with Treatment for both Diseases. International Journal of Mathematical Sciences and Optimization: Theory and Applications. Vol. 8, No. 1, pp. 49-73

[14] Makridakis, S.,and Hibon, M. (1997). ARMA models and the Box-Jenkins methodology. Journal of Forecasting, 16(3), 147-163.

[15] Taiwo A. I, Olatayo T. O, Adedotun A. F. and Adesanya K. K (2019): Modeling and Forecasting Periodic Time Series data with Fourier Autoregressive Model. Iraqi Journal of Science, 2019, Vol. 60, No.6, pp: 1367-1373.

[16] Onasanya O. K, Olusegun A. O., Adedotun A. F. and Odekina G. O. (2021): Forecasting of Exchange Rate in Regime Switching: Evidence from Non ?Linear Time Series Model. The International Journal of Engineering and Science (IJES). 10(09): pp. 01-12

[17] Martínez F, Frías MP, Pérez-Godoy MD, Rivera AJ. (2018). Dealing with seasonality by narrowing the training set in time series forecasting with kNN. Expert Systems with Applications 103:38?48 DOI 10.1016/j.eswa.2018.03.005.

[18] Adesina, O.S., Adedotun, A.F., Oladepo, D.S., Adesina, T.F. (2022): Knowledge, attitude, and perception of health and non-healthcare workers towards COVID-19 vaccination: Machine learning approach. International Journal of Sustainable Development and Planning, Vol. 17, No. 7, pp. 2015-2021. https://doi.org/10.18280/ijsdp.170702