I.J. Information Technology and Computer Science, **2025, 5, 68-83**
Published Online on October 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2025.05.06

A Comparative Study of Statistical (SARIMA) Vis-À-Vis Some Traditional Machine-Learning and Deep-Learning Techniques to Forecast Malaria Incidences in Kolkata of India

Krishnendra Sankar Ganguly
Ernst & Young GDS, Kolkata, India
E-mail: krish.ksg@gmail.com
ORCID iD: https://orcid.org/0000-0002-8259-1419

Krishna Sankar Ganguly
The Kolkata Municipal Corporation, Kolkata, India
E-mail: dr.k.s.ganguly@gmail.com
ORCID iD: https://orcid.org/0000-0001-7470-8413

Ambar Dutta*
Amity Institute of Information Technology, Amity University, Kolkata, India
E-mail: ambardutta@gmail.com
ORCID iD: https://orcid.org/0000-0001-6828-4044
*Corresponding author

Received: 25 April 2025; Revised: 19 June 2025; Accepted: 14 August 2025; Published: 08 October 2025

Abstract: To augment the accuracy of the results of a Time-Series Forecasting problem in the Computational Epidemiology domain of Public Health, to generate an accurate alert in a Real-time Outbreak and Disease Surveillance (RODS) system, namely in the prediction of Malaria incidences, an interdisciplinary approach of data analysis [through Statistical along with Machine-Learning (ML) and Deep-Learning techniques (DL)] has been studied in this research. Two different Non-linear Deep-Learning based techniques, viz., Long Short-Term Memory (LSTM) [a subclass of Recurrent Neural Network (RNN)] & Gated Recurrent Unit (GRU) and two different Non-linear Machine-Learning techniques, viz., Random Forest Regressor & Non-linear Support Vector Machine Regressor are applied in this study to compare against the traditional Statistical-based linear SARIMA model, to forecast a longitudinal data-set of malaria incidences. While SARIMA or other traditional Autoregressive (AR) models, necessitating a smaller number of parameters, undergo limited training and limited prediction power, ML and DL models show profound and persistent performance improvement with better noise-handling/ missing values and perform multi-step forecasts. Moreover, the over-fitting issue can be combated by introducing densely connected residual links in the ML/ DL networks.

Index Terms: Computational Epidemiology, Public Health, Machine Learning, Deep Learning, Real-time Outbreak and Disease Surveillance System, LSTM, GRU, SARIMA, RF Regressor, Non-linear SVM Regressor.

1. Introduction

Big data is being produced at an unprecedented pace. Sorting a large volume of data or clustering it according to public health, a vital domain within health sciences, necessitates periodic prediction of disease epidemics and outbreaks. Timely detection, monitoring, and forecasting of disease outbreaks are quite crucial for implementing effective health interventions to control epidemics. Computational epidemiology primarily focuses on compartmental models that divide the total population into sub-groups of susceptible, infective, and recovered individuals, with transitions among these groups modelled by differential equations. However, traditional (statistical) models, despite their popularity, exhibit limited predictive power due to their inability to model individual-level information and their constrained functional space. These models do not incorporate modern technologies such as machine learning (ML) and data mining. Therefore,

advancing research to develop more resilient predictive models that accommodate the complexities inherent in disease dynamics is imperative to ensure the continued relevance of computational epidemiology in public health management.

An interdisciplinary approach to the spatiotemporal characteristics inherent in time-series forecasting problems requiring real-time alerts reveals that epidemic observations are typically encompassed by Autoregressive (AR) models and their variants (e.g., VAR) to capture temporal patterns [1-3]. While AR models leverage continuously updated historical data to generate linear predictions based on model parameters, Gaussian Process Regression (GPR) [4] enhances predictive capabilities through the use of a non-linear kernel (e.g., radial basis function) to model intricate temporal patterns. Traditional compartment models maintain fixed parameters throughout the process, whereas time-series models exhibit adaptability. Both AR and GPR, due to their reliance on linear combinations or predefined kernels, demand fewer parameters. However, weekly sampled epidemic data undergo limited training, which constrains the expressiveness and predictive power of these widely used models.

Adapting advanced methodologies from other domains to computational epidemiology, the deep learning (DL) approach, specifically deep neural networks, has been applied to the epidemic prediction problem from a time-series forecasting perspective [5-12]. Utilizing adjacent graph convolution and a recurrent module, this method demonstrates significant and consistent performance improvements over baseline methods when tested on multiple real-world datasets [13-17]. Additionally, ablation tests reveal that the overfitting issue, commonly encountered in DL, has been effectively mitigated through the incorporation of densely connected residual links within the networks.

A neural network captures the dynamics of sample data systems in predicting the behavior of a complex, non-linear system to forecast time series predictions of malaria incidences. Recurrent Neural Network (RNN), in particular, a type of DL technique for its capacity to gather insights from sequences [5,18,19], is thus quite suitable to get insight for time-series forecasting [20-22]. The Long Short-Term Memory (LSTM), a subclass of RNN [23-25] enables real-time training of sequential data of any duration, using recurrent, hidden & gated units to notice hidden long/short-term sequential dependencies. Accordingly, RNN is utilized to identify the data structure and pattern, to capture non-linearity and complexity. Thus, a Malaria Incidence Prediction system using RNN-LSTM is envisaged in this paper.

Gated Recurrent Unit (GRU), another DL technique, is also an important sequential data modelling method for its ability to capture long-range dependencies. GRU is conspicuous for its simplified architecture, utilizing lesser numbers of parameters compared to LSTM. Thus, GRU can be used for more efficient training and reduced computational necessities.

The non-linear ML model, Random Forest Regressor/ Random Forest (RF), a versatile group-learning method, is also capable of handling complex data relationships through a combination of multiple decision trees. Random Forest is excellent in capturing non-linear patterns and interactions within the dataset. Another non-linear ML model, Non-linear Support Vector Machine Regressor/ Non-linear Support Vector Machine (SVM) is a powerful regression algorithm, used to handle non-linear relationships effectively. SVM constructs a hyperplane in a high-dimensional space to classify data points, and capture intricate data patterns. Accordingly, Random Forest and SVM are also used in this paper to predict Malaria incidence time series data.

1.1. Objectives of the Present Study

The present research at the initiation used the traditional Statistical (SARIMA) model for forecasting malaria incidences in Kolkata (India). With the gradual development of Machine-Learning and Deep-Learning approaches in malaria predictions, traditional ML models (RF and SVM) and traditional DL models (LSTM & GRU) algorithms were next tried in this research. This paper thus envisages a comparison amongst the above-stated models only, to provide a primary baseline of comparison among traditional, well-documented and/or robust statistical, ML and DL techniques. However, the potentials of the other renowned models like XGBoost for handling high-dimensional datasets, ARIMA-ANN hybrids for combining linear and non-linear dependencies, and Transformers for their capability in long-range sequence modelling are kept reserved for future analysis/ detailing and fine-tuning of this research to provide deeper and more general and comprehensive insights.

Malaria outbreak prediction is a mandatory activity of public health authorities for preventive measures in India. ARIMA, a commonly used linear forecasting model, exhibits vulnerability in outlier detection. When outliers are present, parameters undergo re-estimation as the order of the ARIMA model changes, limiting prediction-adaptability. To improve modelling, various activities are considered, e.g., incorporating trend and seasonality in ARIMA to augment accuracy and address missing values in traditional methods. On the contrary, ML and DL models are advantageous in time-series forecasting due to their capability in accommodating multivariate inputs for better approximating arbitrary nonlinear functions through multi-step predictions, mitigating noise/ missing values. Thus, ML /DL techniques utilize well-defined malaria outbreak parameters [26] for effective prediction [27], while traditional methods rely on in-depth statistics.

As discussed above, some of these diverse ML, DL and traditional ARIMA models are studied in this research, to get an overview of a comprehensive and robust analysis of Malaria incidence forecasting to understand the strengths of traditional Statistical Models against the DL and ML approaches. The objective of this study is to examine the comparative results of malaria prediction obtained from different Statistical and ML or DL algorithms.

1.2. Contributions of the Present Work

The main contributions of this work are:

- A novel comparative effectiveness of two modern ML (Random Forest and SVM) and DL (LSTM & GRU) algorithms against traditional Statistical ARIMA algorithms.
- The limitations of traditional Statistical algorithms in the present era of Neural Net and Artificial Intelligence.
- A significant interdisciplinary application in computational epidemiology in the domain of public health management.
- Forecasting capabilities of ML & DL models in real-time time-series observational data-set of spatiotemporal characteristics.

1.3. Organization of the Paper

The remaining part of the paper is structured as follows. Section 2 deals with a brief review of the related works of literature. Section 3 details the methodologies Data-Collection, followed to arrive at the results for comparison of the five models selected in this study. Section 4 is dedicated to describing the outcomes of the implementation of one statistical, two deep learning and two machine learning models. In the concluding Discussions Section, the prospective future research works are presented followed by conclusions and future works.

2. Related Research Works

For epidemiological forecasting of disease incidences, deaths, prevalence, transmission, hospital resources, etc., various artificial intelligence models (ML/DL) are being widely used even in pandemics (including COVID-19), to medical diagnosis, new drug development, vaccine discovery, disease-tracking, disease-monitoring and time-series of the epidemic [28-29]. Due to factors, viz., lack of training data, heterogeneity of data sources, need for interdisciplinary cooperation, (like, computer science, mathematics, statistics, biology, virology, pharmaceuticals, medicine, etc.), [30], etc., ML & DL models, for their inherent methodological qualities (e.g., data-oriented approach, requiring no specialized/ expert knowledge and significant accuracy (cf. mathematical/ statistical models) [31-32] were more utilized for epidemiological predictions in different countries [33-41], in particular DL algorithms generating deterministic projections [42-44].

In respect of time series modelling, to meet different goals and challenges to achieve the objective of accurate & reliable forecasting, probabilistic predictions were so long better used to project incidences/ deaths [45], due to the existence of a high degree of uncertainty in the data. During COVID-19, the dependency relationship amongst various disease variables (e.g., the number of daily deaths having a high correlation with the vaccination status) [46-47] was not taken into consideration in different DL models applied to the pandemic context around the world and that ultimately favorably changed the dynamics of the pandemic [48]. In time-series analysis, the number of reported cases is associated with variables related to seasonality, weather periods, and even length of holidays [49]. Thus, ML/ DL models and their hyper-parameters often represent a significant additional difficulty [50]. There is another challenge of forecasting the number of new cases, considering volatile fluctuations in daily numbers, suggesting analysis of longer periods of input data to build models for reasonable predictions [51].

3. Methodology

3.1. Study Area

This study is done in Kolkata, a metropolitan city of India, situated at 22.5726° N, 88.3639° E, having an area of 1,886.67 km² and a population of approx. 4,500,000 according to the 2011 Census. The administrative structure of The Kolkata Municipal Corporation (KMC) comprises 144 wards within 15 Boroughs. From the environmental point of view, Kolkata has prolonged summer and rains (annual rainfall of about 1850 mm (73 in) occurring between June to September [52]) with a tropical humid, hot, wet-and-dry climate, characterized by an annual mean temperature of 26.8°C (variations ranging from 30° –40°C), while the winter season spanning only 2-3 months, with minimum temperatures dropping to 9–11 °C in December and January [53].

3.2. Data Description

Through the Integrated Disease Surveillance Programme (IDSP) under National Health Mission, public health organizations systematically collect patient-related information and maintain health records in all States and Union Territories of India, which are digitized to a centralized database for efficient data storage and retrieval. Consequently, a substantial volume of data undergoes comprehensive analyses to generate decisions pertaining to disease surveillance and public health management [54-56].

The Kolkata Municipal Corporation, the governing body responsible for managing endemic disease outbreaks in Kolkata, actively gathers, stores, and maintains disease-related data for the city. The dataset comprises malaria case-details recorded by KMC. The records of data, spanning from January 1, 2014, to December 31, 2018 are considered here for analysis. The Malaria Incidence Rate is computed by multiplying the total positive cases per week by 100,000 and dividing the result by the population of Kolkata. This calculated Incidence Rate (reflecting the number of infected individuals in a given week) indicates the risk of infection, which serves as a pivotal metric for the study.

3.3. Data Analysis Methods

The epidemic prediction problem constitutes a time series forecasting task, as evidenced by $X_i \in E^m$, the multivariate epidemiological profile of a disease. This profile comprises observations (x) from various sources/signals at any time point (t), such as the weekly malaria patient counts in m boroughs of Kolkata, influenced by different environmental variables. In this study, a univariate study-data of Malaria Incidence Rate in Kolkata only is considered for analysis. However, this study data is analyzed by five analytical methods to detect optimal analytical methodology. Given $X = [x_1, x_2, \ldots, x_T]$, representing the available training data over a time span of size T, the objective is to forecast the epidemiological profile of malaria in Kolkata at a future time point (T+h), where h denotes the time lag.

Thus, this study uses Time-series analysis using traditional Statistical Auto-Regressive models along with two Machine-Leaning models as well as two Deep-Learning models to forecast weekly malaria incidence rates in Kolkata. Accordingly, this paper aims to compare the results obtained from Machine-Leaning models and Deep-Learning models algorithms against the results of traditional Statistical Auto-Regressive models to forecast time-series data.

3.4. Performance Evaluation of Different Models

Root Mean Squared Error (RMSE) is a common measurement for the accuracy of the prediction. Here, the RMSE value is used to compare the performance of the prediction of the five models used here.

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{1}$$

The Root Mean Squared Error (RMSE) serves as the metric for assessing model performance, measuring the deviation between observed and predicted values. A lower RMSE indicates reduced variance between actual and predicted results, reflecting improved model accuracy in predictive tasks.

Mean Absolute Error (MAE) measures the average magnitude of errors between predicted and actual values, irrespective of their direction. It provides an absolute scale of prediction accuracy. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}| \tag{2}$$

Mean Absolute Percentage Error (MAPE) quantifies prediction accuracy as a percentage. It calculates the average absolute percentage difference between actual and predicted values:

$$MAPE = \frac{1}{n}\left|\frac{X_i - \bar{X}}{X_i}\right| \times 100 \tag{3}$$

4. Results

Based on the observed weekly data of Malaria Incidence Rates of Kolkata from Jan-2014 to Dec-2018, a plot is drawn, where $Y_1, Y_2, \ldots, Y_N$ show the length N (N = 1, 2, 3, …..261) data points (Fig.1.). In this time-series plot, a seasonal pattern of a peak in July to September is prominent along with a small peak in November-December. As is observed, the pattern of Malaria Incidence Rates is downward in 2015 and from 2016 to 2018 it is almost the same.

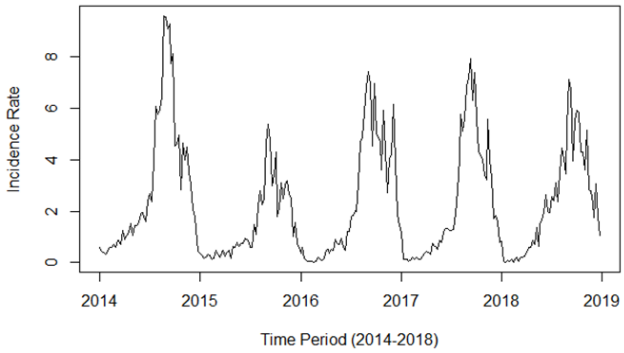


Fig.1. Time series plot of malaria incidence rate

4.1. Data Pre-processing

While feature scaling and normalization are often recommended for ML models like SVM and RF, for data distance measurements, GRU and LSTM focus on sequence learning rather than absolute distances between values. GRU and LSTM models have inherent learning capabilities and are designed to handle sequential data effectively and to focus on learning patterns over time. As the models use different gates (viz., input, forget, and output gates) to control the flow of

information, issues related to the scale or magnitude of the input data are mitigated. The ability to focus on relative changes and temporal dependencies reduces the sensitivity of these models to the absolute scale of the input values. GRU and LSTM models are resilient to input scaling and they internally learn weights and biases during training that adapt to the scale of the data. As long as the input values are consistent across training and testing phases, these models perform well without explicit scaling or normalization. In the present study, for uni-variate forecasting tasks, the input typically represents a single feature (e.g., malaria incidence counts), making scaling less critical compared to multivariate tasks where features have varying scales.

However, employing imputation techniques viz., forward filling, missing values were handled before feeding the data into GRU and LSTM models, to ensure that the sequence remained continuous and meaningful for temporal modelling.

A. Time-series Stationarity

Using the moving averages technique, the Time Series (Y_t) is decomposed into seasonality (S_t), trend (T_t) & irregular (e_t) components [57] which reveals a downward trend from 2014 to 2015 and then the trend from 2016 to 2018 become same, having seasonality of 52 weeks period. The decomposition of the time series depicting trends and 52-week seasonality patterns is shown in Fig.2.

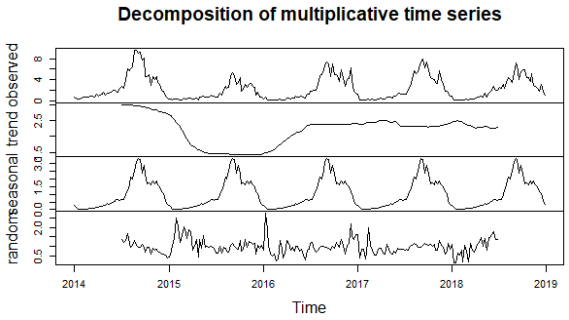


Fig.2. Decomposition of time series plot of malaria incidence rate

Upon determining that the time series is non-stationary via the Augmented Dickey-Fuller (ADF) test (Table 1.), the trend and seasonal components were eliminated by applying s[th]-order differencing (where s=52) difference ($\nabla_{52} Y_t = \widetilde{Y}_t$) of the time series. Post-transformation, the ADF test verified the stationarity of the time series (Table 2.).

Table 1. ADF test result (Pre-transformation)

Augmented Dickey-Fuller (ADF) Test Dickey-Fuller = -3.143, Lag order = 0, p-value = 0.09783 Alternative hypothesis: stationary

Table 2. ADF test result (Post-transformation)

Augmented Dickey-Fuller Test lag = 52 Dickey-Fuller = -5.8828, Lag order = 0, p-value = 0.01 alternative hypothesis: stationary

B. Data Splitting

Before model fitting, the time-series dataset is divided into two parts viz. Training Part and Testing Part. The dataset is partitioned into 75% training and 25% testing subsets, facilitating comprehensive model training and evaluation processes. It facilitates the evaluation of the model's predictive capabilities on unseen data.

4.2. Model Fitting

A. SARIMA Model

Once the time series has been made stationary, the difference parameters (d) and (D) for the SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model are selected, as has already been done by the present authors in several previous studies [58-59]. The seasonal parameters (P) and (Q) are then identified based on the patterns observed in the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) at various lags.

Using model selection criteria, viz., the Akaike information criterion (AIC), it is found that (1, 1, 2) as (p, d, q) and (0, 1, 0) as (P, D, Q) come out to be the best combinations vide Table 3.

Among the various potential SARIMA models evaluated for the Malaria Incidence Rate time series, the model with the lowest selection criterion value was determined to be optimal [60-61]. Subsequent iterations revealed the presence of one non-seasonal autoregressive term (AR) and two non-seasonal moving average terms (MA), corresponding to p=1 and

q=2, respectively. The parameters are determined by iterations with least AIC and BIC.

Table 3. Best SARIMA model

ARIMA (0,1,3)(1,1,1)[52]	: Inf
ARIMA (0,1,2)(1,1,1)[52]	: Inf
.	
ARIMA (0,1,0)(0,1,0)[52]	: 475.7571
.	
ARIMA (1,1,2)(1,1,1)[52]	: Inf
ARIMA (1,1,0)(0,1,0)[52]	: 460.4429
ARIMA (1,1,2)(0,1,0)[52]	: 455.3975
Best model: ARIMA (1,1,2)(0,1,0)[52]	

The forecasted values of the Incidence Rate data series are given in Fig.3., which clearly shows observed Incidence Rate data series for the year 2018 lies within 80% confidence intervals and is close enough to the point forecast for the year 2018.

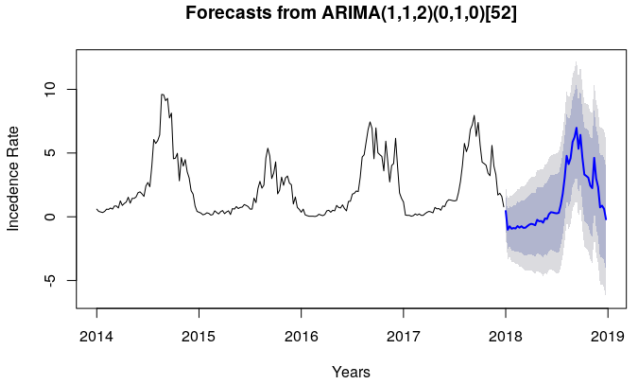


Fig.3. Time series plot (Training and predicted data)

B. Random Forest Model

Utilizing the ScikitLearn library in Python, a thorough exploration of the Random Forest model is conducted to analyze the Malaria Incidence Rate dataset from January 2014 to December 2018, with strict adherence to rigorous methodology and evaluation techniques. Moreover, hyperparameter optimization plays a pivotal role in enhancing model efficacy. Through the application of the GridSearchCV technique, multiple hyper-parameters (Table 4.) are rigorously fine-tuned, ensuring the model's optimal configuration.

Table 4. Hyperparameters used in random forest model

Hyperparameters	Values
n_estimators	(10, 17, 25, 33, 41, 48, 56, 64, 72, 80)
max_features	('auto', 'sqrt', 'log2')
max_depth	(2, 4, None)
min_samples_split	(2.5)
min_samples_leaf	(1,2)
bootstrap	(True, False)

The best estimator across all searched parameters in the Random Forest Model is given in Table 5.

Table 5. Best estimators used for random forest model

RandomForestRegressor ('bootstrap': True, 'max_depth': 2, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 1000)

The fitted Random Forest Model Architecture is given in Fig.4.

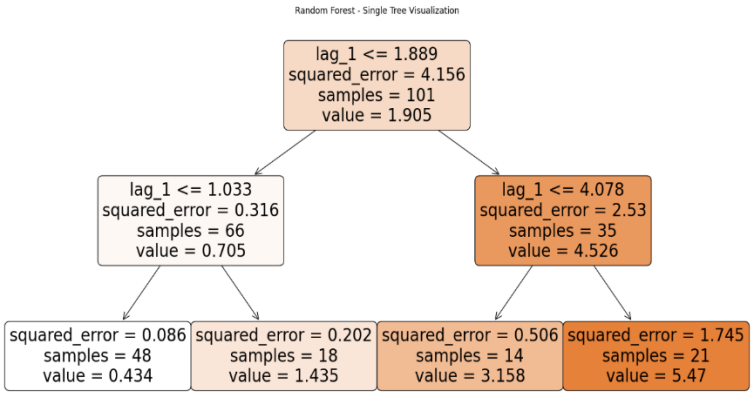


Fig.4. Fitted random forest model tree architecture

The fitted Plot of the RF model (Fig.5.) is given below.

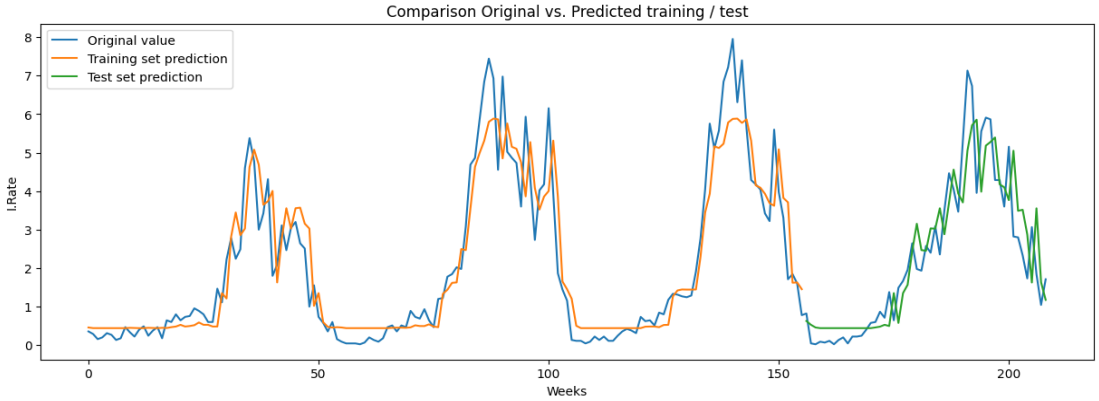


Fig.5. Original and predicted data plot after fitting random forest model

C. Support Vector Machine Model

The implementation of the SVM Regressor model over the Malaria Incidence Rate dataset is characterized by several key steps, each contributing to the understanding of the model's performance. After data pre-processing, the Support Vector Machine Regressor model is instantiated using Python's Scikit-Learn library, ensuring robust and efficient model development. Then, to optimize the model's hyperparameters, the GridSearchCV technique is leveraged, to systematically explore a range of values for key parameters such as kernel types, gamma, epsilon, etc. (Table 6.).

Table 6. Hyperparameters used in support vector machine regressor model

Hyperparameters	Values
kernel	('poly','sigmoid','rbf')
C	(0.01,0.1,1,10)
gamma	(0.01,0.1,1)
epsilon	(0.01,0.1,1)
shrinking	(True, False)

This approach enables the identification of the optimal combination of hyperparameters that maximizes the performance of the model while minimizing computational overhead. The best estimator used across all searched parameters in SVM Regressor Model are given below in Table 7.

Table 7. Best Estimators for support vector machine regressor model

SVR('C': 10, 'epsilon': 0.1, 'gamma': 0.01, 'kernel': 'rbf', 'shrinking': True)

The fitted time series model (Fig.6.) is given below.

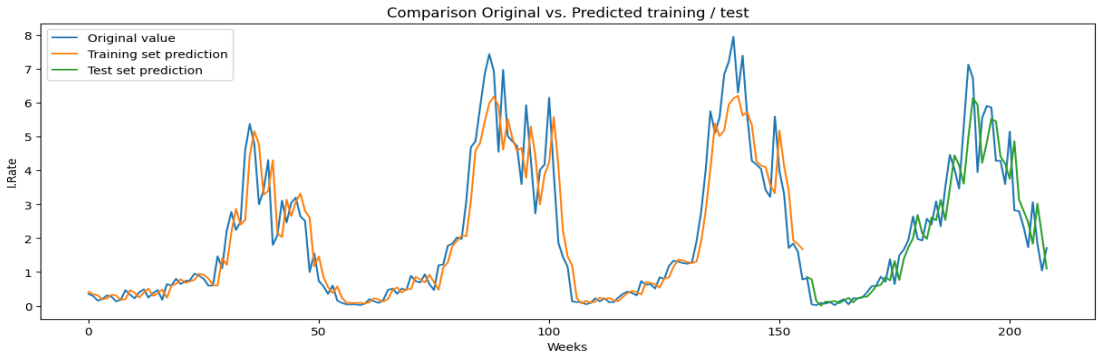


Fig.6. Original and predicted data plot after fitting support vector machine regressor model

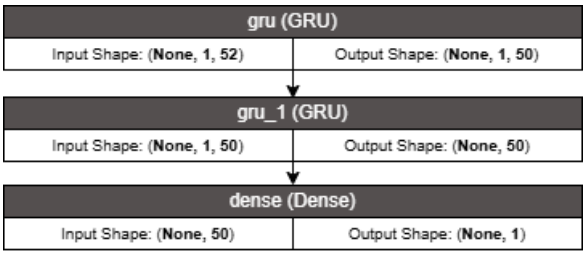


Fig.7. Fitted GRU model architecture

Table 8. GRU model training

```
Epoch 1/50
5/5 - 4s - 733ms/step - loss: 0.1246 - val_loss: 0.0635
Epoch 2/50
5/5 - 0s - 11ms/step - loss: 0.0537 - val_loss: 0.0248
Epoch 3/50
5/5 - 0s - 11ms/step - loss: 0.0386 - val_loss: 0.0178
Epoch 4/50
5/5 - 0s - 11ms/step - loss: 0.0270 - val_loss: 0.0132
.
.
.
Epoch 47/50
5/5 - 0s - 11ms/step - loss: 0.0066 - val_loss: 0.0079
Epoch 48/50
5/5 - 0s - 11ms/step - loss: 0.0065 - val_loss: 0.0077
Epoch 49/50
5/5 - 0s - 11ms/step - loss: 0.0064 - val_loss: 0.0076
Epoch 50/50
5/5 - 0s - 11ms/step - loss: 0.0067 - val_loss: 0.0081
7/7 [==============================] - 1s 2ms/step
3/3 [==============================] - 0s 3ms/step
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
gru (GRU)                   (None, 1, 50)             15600

gru_1 (GRU)                 (None, 50)                15300

dense (Dense)               (None, 1)                 51
=================================================================
Total params: 92855 (362.72 KB)
Trainable params: 30951 (120.90 KB)
Non-trainable params: 0 (0.00 Byte)
Optimizer params: 61,904 (241.82 KB)
```

D. GRU model

The GRU (Gated Recurrent Unit) model, employed in this study using the Keras library, consists of two GRU layers, each with 50 neurons, followed by a Dense layer with a single neuron to output the predicted value (Fig.7.).

The model was optimized utilizing the Adam algorithm, with the training process guided by the Mean Squared Error (MSE) loss function applied to the training dataset. The training involved 50 epochs with a batch size of 32. During training, the model's performance was monitored on both the training and validation sets to ensure it was learning

effectively without overfitting (Table 8.). To assess the generalizability of the model, its performance on the validation set was monitored throughout the training process. The training loss, measured by MSE, decreased steadily over the epochs, indicating that the model has been effectively learning from the training data. The validation loss followed a similar decreasing trend, suggesting that the model was not overfitting and was generalizing well to the validation data.

The fitted GRU model based on the Malaria Incidence Rate is shown below in Fig.8.:

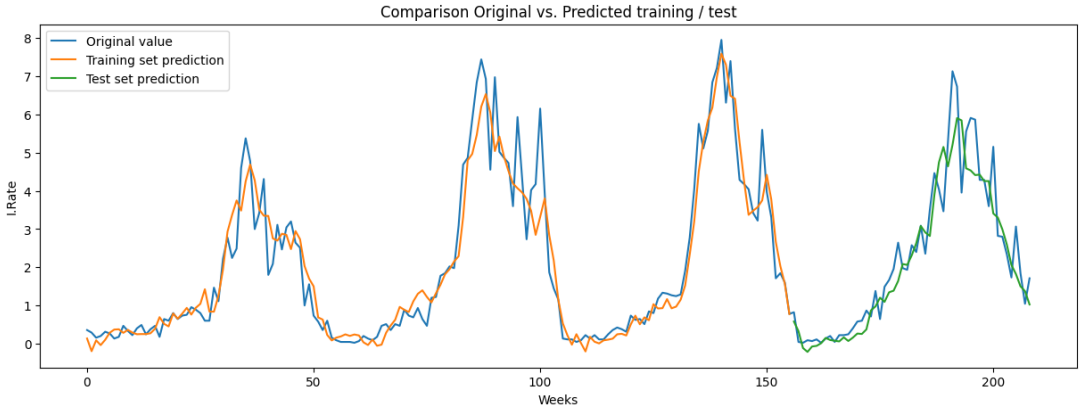


Fig.8. Original and predicted data plot after fitting GRU model

E. RNN-LSTM Model

In this study, The RNN-LSTM model is employed using the Keras library. Among various parameters of the LSTM layer, values of some parameters, viz. units, activation functions, dropouts and return-sequence are modified while other parameters use the default values.

The performance of the Adam and Stochastic Gradient Descent (SGD) optimizers is compared, utilizing their default parameter settings. This study also examines the effects of various activation functions, including Rectified Linear Unit (ReLU), Sigmoid, Exponential Linear Unit (ELU), and Hyperbolic Tangent (tanh), to evaluate and compare outcomes. In addition, various epochs (100, 200, 500, and 1000) were used to evaluate results. Here, four hidden layers are employed, with each hidden layer consisting of 50 units. The value of the batch size used here is 50 while the drop-out value of each hidden layer used here is 0.2. The models are built using the variable and fixed parameters, following training on the training data (Fig.9.). The generated models are then evaluated to identify the most accurate predictive model.

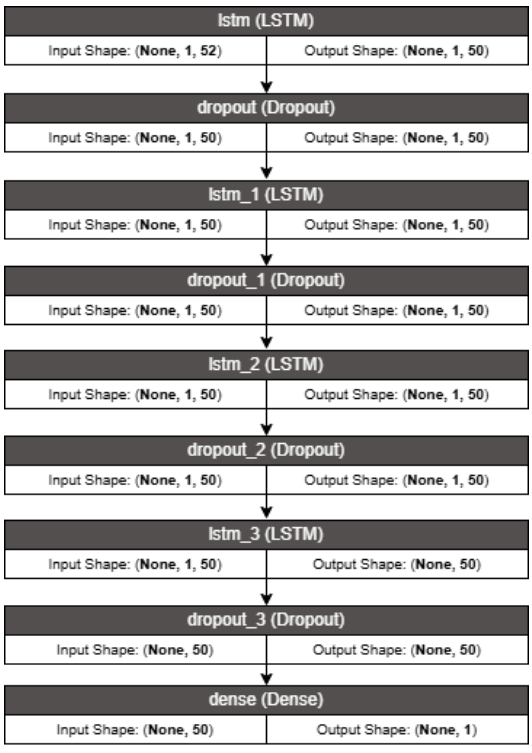


Fig.9. Fitted LSTM model architecture

The fitted RNN-LSTM model over Malaria Incidence Rate is shown below in Fig.10.:

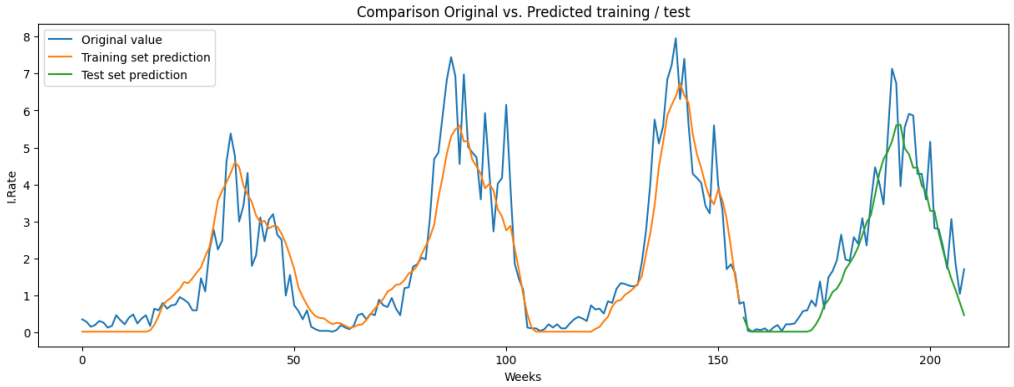


Fig.10. Original and predicted data plot after fitting RNN-LSTM model

4.3. Forecasted Data

The predicted data (Table 9.) for each model along with the original incidence rate for 52 weeks of the year 2018 are given below.

Table 9. Original and predicted data values after fitting models

WEEK	ORIGINAL INCIDENCE RATE	SARIMA	RF REGRESSOR	SVM REGRESSOR	GRU	LSTM
1	0.044444444	0.4508784	0.527075546	0.779732011	0.328191811	0.114397079
2	0.022222222	1.0171834	0.45772244	0.143274286	-0.114760102	0.022222222
3	0.088888889	0.7493818	0.441066095	0.006572284	-0.216621612	0.022222222
4	0.066666667	0.9380457	0.441066095	0.124750341	-0.07504714	0.022222222
5	0.111111111	0.8819843	0.441066095	0.11817158	-0.058684393	0.022222222
6	0.022222222	0.9173759	0.441066095	0.14577872	0.010445283	0.022222222
7	0.133333333	0.7321083	0.441066095	0.08022159	0.149006122	0.022222222
8	0.2	-0.854781	0.441066095	0.155786855	0.094785914	0.022222222
9	0.044444444	0.7439153	0.441066095	0.236795703	0.074980687	0.022222222
10	0.222222222	0.8693226	0.441066095	0.105573854	0.058076943	0.022222222
11	0.222222222	0.8600228	0.441066095	0.226281344	0.16400222	0.022222222
12	0.244444444	0.7327391	0.441066095	0.267067159	0.069296509	0.022222222
13	0.4	0.6176926	0.441066095	0.270186489	0.156549507	0.022222222
14	0.577777778	0.5535859	0.441066095	0.407747865	0.264422501	0.022222222
15	0.6	0.5963651	0.459054661	0.586165002	0.253659211	0.022222222
16	0.866666667	0.6641151	0.479597895	0.621904361	0.373344306	0.064566175
17	0.711111111	0.2411881	0.527907913	0.853129401	0.87076256	0.212468085
18	1.377777778	0.3527577	0.495887507	0.752703716	0.97423621	0.41666753
19	0.644444444	0.3302372	1.351424255	1.326278662	1.202840657	0.731004734
20	1.488888889	0.4637646	0.576711553	0.763221046	1.094233011	0.880064554
21	1.666666667	-0.130305	1.351424255	1.391602504	1.344968889	1.095506196
22	1.955555556	0.1748316	1.566725818	1.735338304	1.386416204	1.188708895
23	2.644444444	0.2029996	2.449291726	1.998550515	1.640029576	1.380930278
24	1.977777778	0.3585204	3.153193516	2.689848459	2.085024289	1.703236414
25	1.933333333	0.3363208	2.468278978	2.138152839	2.061543541	1.873591643
26	2.577777778	0.2918617	2.451593832	1.978103171	2.311598261	2.063773566
27	2.4	0.269649	3.033490979	2.611675821	2.631535587	2.306096494
28	3.088888889	0.3140872	3.024337894	2.533334716	3.080622343	2.631534405
29	2.355555556	0.9363135	3.556880725	3.132287363	2.910999107	2.973759207
30	3.511111111	1.8029775	2.880925524	2.539072642	2.819595062	3.172055468

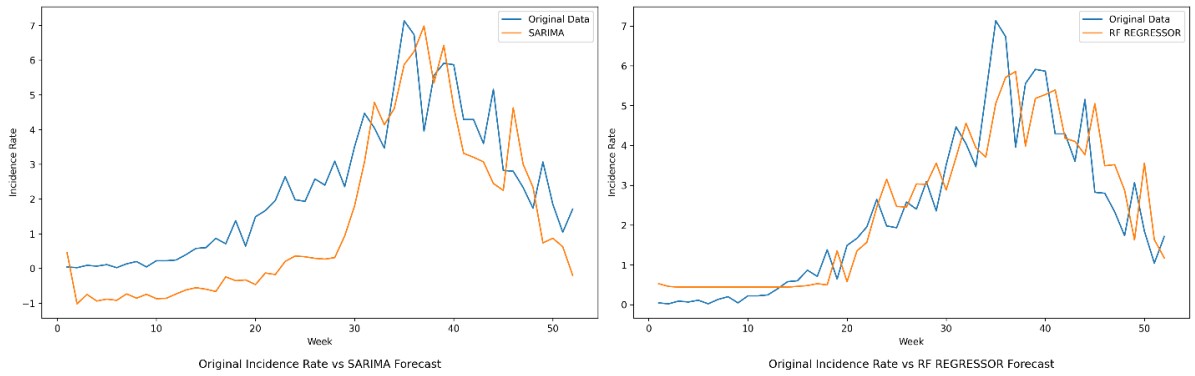
31	4.466666667	3.0696459	3.703238276	3.479990303	3.895774024	3.714281678
32	4.044444444	4.7807559	4.559385315	4.449032489	4.746511286	4.304604254
33	3.466666667	4.1363122	3.939059363	4.157996797	5.151571374	4.6868936
34	5.288888889	4.6029784	3.707244775	3.607318149	4.64227278	4.892932055
35	7.133333333	5.8696454	5.052003431	4.964259087	5.20338965	5.171258574
36	6.733333333	6.2474229	5.712630027	6.13959905	5.906517364	5.601425824
37	3.955555556	6.9807564	5.85760585	5.949500408	5.846286847	5.617376456
38	5.555555556	5.3363119	3.986350185	4.227203983	4.591715158	4.97858797
39	5.911111111	6.4252008	5.183584958	4.817301599	4.536455855	4.819685496
40	5.866666667	4.647423	5.277760943	5.532639806	4.414338379	4.454423492
41	4.288888889	3.3140897	5.39500358	5.455382446	4.42691182	4.460083196
42	4.288888889	3.2029786	4.178172713	4.409483896	4.256921642	3.978282324
43	3.6	3.0696452	4.100696179	4.204770371	4.257112206	3.805651203
44	5.155555556	2.447423	3.76645054	3.754374294	3.409540766	3.28824158
45	2.822222222	2.247423	5.051111733	4.870561461	3.301590754	3.2845549
46	2.8	4.6252008	3.489091445	3.144105044	2.994418142	2.699789889
47	2.333333333	3.0029786	3.517170336	2.81264788	2.591094644	2.234447268
48	1.733333333	2.3363119	2.874806577	2.471865657	2.061145154	1.811628081
49	3.066666667	0.7363119	1.625876925	1.836303404	1.822086876	1.448777208
50	1.844444444	0.8696452	3.554578619	3.019711336	1.501400297	1.146489512
51	1.044444444	0.6252008	1.632883412	2.066050789	1.373665203	0.809170669
52	1.711111111	0.1970214	1.175234427	1.105297408	1.025837795	0.469559501

A consolidated chart showing Original Incidence Rate against forecasted data of each model is given below in Fig.11.

4.4. Residual Analysis

To determine any significant difference existing between the means of two groups of model performance-metrics across models and to know how they are related, initially Student's t-test (Paired t-test) and the one-way analysis of variance (ANOVA) test may be conducted. But, t-Test and ANOVA results are insufficient to identify the best-performing model due to their reliance on mean differences, failing to capture the variability or distributional differences in errors of the models and the statistical assumptions (particularly normality and equal variances). To explore which specific group-mean of error distributions in time series forecasting is different from which group, a post-hoc test may be done for some reliable conclusions, though it's not so simple because data may not meet the distributional assumptions of ANOVA. In such case, a non-parametric test is performed, as an analog to one-way ANOVA. Non-parametric test allows to completely drop normality assumptions about the residual distribution. Instead of testing whether the means of groups are equal or not, it tests whether the entire distribution of values are generally in the same location or not.

The Kolmogorov-Smirnov (K-S) test, a non-parametric statistical method, is performed to assess whether the residual series conforms to a specified distribution [$r_k : N(0, \frac{1}{N})$] under the null hypothesis. This test is robust and does not assume a specific distribution for the data. The null hypothesis in the K-S test is accepted if the test statistic is below the critical value, indicating that the residuals conform to the hypothesized distribution. Conversely, the null hypothesis is rejected if the test statistic exceeds the critical value, suggesting a deviation from the specified distribution [62-63].



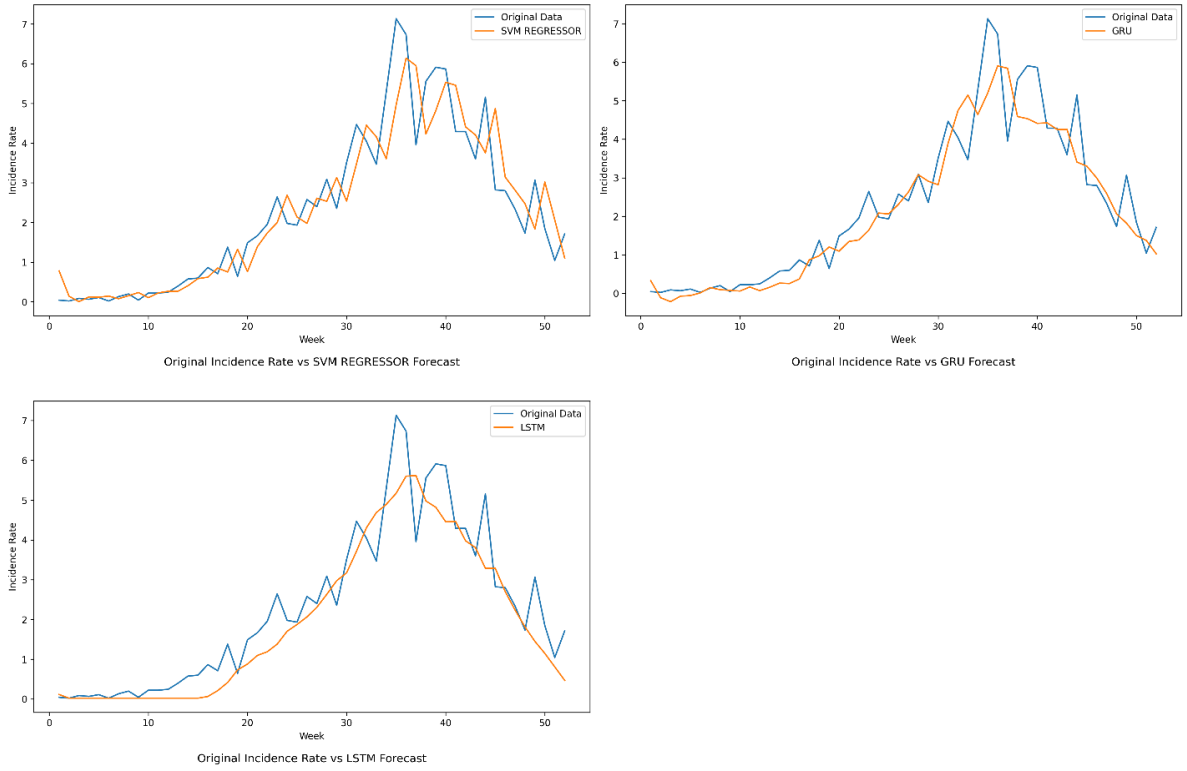


Fig.11. Original and predicted data plot of each model

The parametric Box-Ljung Test is applied to evaluate the null hypothesis that residual data values are random and independent up to a predetermined number of lags. This test is widely utilized for detecting autocorrelation in time series data. The null hypothesis is accepted if the p-value exceeds a selected significance level, usually set at 0.05, suggesting the absence of significant autocorrelation in the residuals. Conversely, if the p-value falls below the significance level, the null hypothesis is rejected, indicating the presence of autocorrelation [64].

The results of both tests for each model are presented in Table 10. below.

Table 10. K-S Test and Box-Ljung test result for all models

Model Name	K-S Test (p-value)	Box-Ljung Test (p-value at lag=48)
Seasonal Auto-Regressive Integrated Moving Average	0.8533	0.6181
Random Forest Regressor	0.8128	0.3375
Support Vector Machine Regressor	0.6795	0.2357
Gated Recurrent Unit	0.5049	0.1334
RNN-Long Short-Term Memory	0.4458	0.4257

Table 10. shows that the p-values of the K-S test for all models are greater than 0.05 and hence at a 5% level of significance the null hypothesis is accepted. Also, for the Box-Ljung test, p-values for all models are greater than 0.05, which suggests that the data values of the residuals are random and independent up to a certain number of lags. Hence, it can be said that the fitted models are adequately describing the Incidence Rate data series, without further search for any modification of the fitted model.

4.5. Performance Evaluation of Models

As mentioned earlier, the performance of the models is evaluated using the Root Mean Squared Error (RMSE) score. RMSE, being a robust measure of prediction error, provides insights into the magnitude of deviations between predicted and actual Malaria Incidence Rates. The obtained RMSE score for each model indicates strong predictive accuracy, where lower values suggest a closer fit between the predicted and actual values of the target variable.

The detailed comparison of RMSE scores of different Models is tabulated vide Table 11.

Table 11. Every model's performance evaluation using RMSE score

Model Name	RMSE Score	MAE Score	MAPE Score
Seasonal Auto-Regressive Integrated Moving Average	1.34	1.26	393.68
Random Forest	0.88	0.71	173.80
Support Vector Machine Regressor	0.81	0.61	90.42
Gated Recurrent Unit	0.71	0.51	64.66
RNN-Long Short Term Memory	0.74	0.53	40.99

Among the models tested, the Deep Learning Models, both GRU and RNN-LSTM model performs admirably with the lowest almost similar RMSE scores of 0.71 and 0.74 respectively, indicating the highest accuracy in predicting future data points. The Support Vector Machine Regressor (SVM) model also performs well, with an RMSE of 0.81. The Random Forest model, with an RMSE score of 0.88, stood fourth in terms of RMSE. The Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model, having the highest RMSE score of 1.34, shows the least accuracy among the models in terms of RMSE. This suggests that SARIMA can explain the general trend and variability of the data but lacks the precision in individual predictions the other models can achieve.

SARIMA performed the worst, with an MAE of 1.26 and a high MAPE of 393.68, indicating its inability to handle the nonlinear complexities of malaria time series data. RF showed improvement with an MAE of 0.71 and MAPE of 173.80, reflecting its capability to manage variability. SVM further reduced errors (MAE: 0.61, MAPE: 90.42), demonstrating better modelling of nonlinear patterns.

DL models outperformed traditional approaches. GRU achieved an MAE of 0.51 and MAPE of 64.66, showcasing its ability to model temporal dependencies effectively. LSTM excelled with the lowest MAE (0.53) and MAPE (40.99), owing to its memory cell structure that captures long-term dependencies.

These results suggest that while traditional methods like SARIMA are still useful, advanced ML and DL models, particularly GRU and RNN-LSTM, may provide more precise and reliable predictions for time-series data. The lower RMSE, MAE & MAPE scores of both models indicate better accuracy in individual predictions. Future research could explore hybrid models and additional metrics to enhance predictive accuracy and reliability.

4.6. Discussions

LSTM incorporates three gates (input, forget, and output) and a memory cell to selectively retain relevant information over long time horizons, allowing the model to overcome the vanishing gradient problem, common in standard RNNs. GRU simplifies the architecture by combining the input and forgetting gates into an update gate while retaining comparable performance (for smaller datasets, like the present research). GRU and LSTM, as specialized Recurrent Neural Network (RNN) architectures, due to their unique gating mechanisms, are designed to handle sequential data and capture long-term temporal (e.g., seasonal trends) dependencies effectively. The superior performance of GRU and LSTM in this study may primarily be attributed to their architectural ability to capture temporal dependencies, to inherently handle noise/ missing/ incomplete values (learning the relationships between sequential data points) and also to model complex non-linear patterns.

In contrast, SARIMA, SVM, and RF are either limited by their reliance on stationary data processing (in comparison to dynamic data processing in DL models), lack of capabilities in sequential modelling and/or face challenges in multi-step forecasting. Moreover, GRU and LSTM can naturally handle multi-step forecasting by learning complex mappings from historical data to future values. They generate predictions iteratively or for multiple steps simultaneously giving them an edge over the traditional SARIMA model (having limitations in multi-step predictions and significant performance degradation). SVM regressor can capture non-linear patterns but is not inherently designed for sequential or temporal data, because SVM treats each data point independently, failing to leverage the sequential data structure effectively. Also, hyper-parameters (e.g., kernel type, C, gamma) and scalability can become a bottleneck for larger datasets for SVM. RF is an ensemble learning method based on decision trees, effective for non-linear relationships, treating each observation as independent without explicit modelling of temporal dependencies in time-series data. RF lacks the ability to extrapolate well, making it less suited for forecasting tasks, especially in capturing trends or seasonality. Thus, in essence, the following attributes drive the superiority of GRU and LSTM over other models, viz., (a) Temporal Feature Extraction, (b) Handling Non-Linear Relationships and (c) Scalability and Flexibility.

5. Conclusions and Future Scope

In conclusion, this research endeavors to advance the field of computational epidemiology by employing a comparative analysis of different time-series forecasting models for malaria incidence rate, particularly focusing on the innovative application of Deep Learning GRU and LSTM networks in contrast to the traditional SARIMA model. The results elucidate the superior predictive performance of the Deep Learning models, as evidenced by consistently lower RMSE values in forecasting weekly malaria incidence rates in Kolkata. Leveraging the capabilities of Deep Learning models, their efficacies have been demonstrated in capturing intricate temporal patterns, offering a promising avenue for enhanced disease forecasting in real-world scenarios. The methodology, integrating variable parameters such as activation

functions, epochs, and window sizes, contributes further to the understanding of the adaptability of the model and robustness in optimizing predictive accuracy. This study outcome not only emphasizes the practical utility of Deep Learning in epidemiological forecasting but also advocates for its integration into public health management strategies for more precise and timely interventions. However, recognizing the dynamic nature of patterns of infectious disease, this study will certainly prompt future investigations into refining GRU and LSTM models, exploring additional parameters, and expanding datasets to bolster the adaptability and generalizability of the concerned Models. Future studies are contemplated with the incorporation of multi-region datasets and broader arrays of techniques to explore the generalizability of the findings.

Time-series prediction usually employs one data property as both input and output, but correlating malaria incidence with temporal and spatial factors requires multiple data series as input. As predicting complex temporal dependences remains challenging even with state-of-the-art Machine Learning and Deep Learning algorithms, Time Series Forecasting is still one of the most intricate problems in real-life scenarios.

References

[1] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," in Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 115, pp. 700–721, 1927.

[2] C. Chatfield, "The Analysis of Time Series: An Introduction," 6th ed., Chapman & Hall, New York, pp. 352, 2003.

[3] K. P. Burnham and D. R. Anderson, "Model selection and multimodel inference: A practical information-theoretic approach," 2nd ed., Springer-Verlag, New York, 2002.

[4] R. Senanayake, S. T. O'Callaghan, and F. Ramos, "Predicting spatio-temporal propagation of seasonal influenza using variational Gaussian process regression," in AAAI, pp. 3901–3907, 2016.

[5] A. Graves, "Supervised sequence labelling with recurrent neural networks," Studies in Computational Intelligence, Springer, 2012.

[6] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," International Journal of Environmental Research and Public Health, vol. 15, no. 8, pp. 1596, 2018.

[7] K. Lee, J. Ray, and C. Safta, "The predictive skill of convolutional neural networks models for disease forecasting," PLOS ONE, vol. 16, no. 7, pp. e0254319, 2021.

[8] S. Dixon, R. Keshava Murthy, D. H. Farber, A. Stevens, K. Pazdernik, and L. E. Charles, "A comparison of infectious disease forecasting methods across locations, diseases and time," Pathogens, vol. 11, no. 2, pp. 185, 2022.

[9] G. S. Bhavekar, A. Das Goswami, C. P. Vasantrao, A. K. Gaikwad, A. V. Zade, and H. Vyawahare, "Heart disease prediction using machine learning, deep learning and optimization techniques - A semantic review," Multimedia Tools and Applications, 2024.

[10] Z. Yu, K. Wang, Z. Wan, S. Xie, and Z. Lv, "Popular deep learning algorithms for disease prediction: A review," Cluster Computing, vol. 26, pp. 1231–1251, 2023.

[11] C. Burghard, "Big data and analytics key to accountable care success," IDC Health Insights, 2012.

[12] G. P. Kanna, S. J. Kumar, Y. Kumar, A. Changela, M. Woźniak, J. Shafi, and M. F. Ijaz, "Advanced deep learning techniques for early disease prediction in cauliflower plants," Scientific Reports, vol. 13, no. 18475, 2023.

[13] W. Raghupathi, "Data mining in health care," in Healthcare Informatics: Improving Efficiency and Productivity, S. Kudyba, Ed., Taylor & Francis, pp. 211–223, 2010.

[14] P. Ravikumaran and K. Vimala Devi, "A review: big data and analytics in health care," Indian Journal of Engineering, vol. 13, no. 31, pp. 1–10, 2016.

[15] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and System, vol. 2, no. 3, pp. 10, 2014.

[16] D. J. Park, M. W. Park, H. Lee, Y. J. Kim, Y. Kim, and Y. H. Park, "Development of machine learning model for diagnostic disease prediction based on laboratory tests," Scientific Reports, vol. 11, no. 1, pp. 7567, 2021.

[17] R. Kaundal, A. S. Kapoor, and G. P. Raghava, "Machine learning techniques in disease forecasting: A case study on rice blast prediction," BMC Bioinformatics, vol. 7, pp. 485, 2006.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[19] I. Sutskever, "Training recurrent neural networks," PhD thesis, University of Toronto, 2013.

[20] E. Busseti, I. Osband, and S. Wong, "Deep learning for time series modeling," Technical report, Stanford University, 2012.

[21] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in Proceedings of the 28th International Conference on Machine Learning, pp. 2011.

[22] W. K. Wong, M. Xia, and W. C. Chu, "Adaptive neural network model for time-series forecasting," European Journal of Operational Research, vol. 207, no. 2, pp. 807–816, 2010.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] Q. Lyu and J. Zhu, "Revisit long short-term memory: An optimization perspective," in Advances in Neural Information Processing Systems Workshop on Deep Learning and Representation Learning, 2014.

[25] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," Neural Computation, vol. 1, no. 2, pp. 270–280, 1989.

[26] R. Sharma, "Epidemiological investigation of malaria outbreak in village Santej, district Gandhi Nagar (Gujarat)," Indian Journal of Preventive and Social Medicine, vol. 37, no. 3&4, 2006.

[27] O. P. Zacarias and H. Boström, "Predicting the incidence of malaria cases in Mozambique using regression trees and forests," International Journal of Computer Science and Electronics Engineering, vol. 1, no. 1, 2013.

[28] F. Kamalov, A. K. Cherukuri, H. Sulieman, F. Thabtah, and A. Hossain, "Machine learning applications for COVID-19: a state-of-the-art review," in Data Science for Genomics, A. K. Tyagi and A. Abraham, Eds., Academic Press, pp. 277–289, 2023.

[29] J. C. Clement, V. Ponnusamy, K. C. Sriharipriya, and R. Nandakumar, "A survey on mathematical, machine learning and deep learning models for COVID-19 transmission and diagnosis," IEEE Reviews in Biomedical Engineering, vol. 15, pp. 325–340, 2022.
[30] C. Comito and C. Pizzuti, "Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: A focused review," Artificial Intelligence in Medicine, vol. 128, pp. 102286, 2022.
[31] J. Devaraj, R. Madurai Elavarasan, R. Pugazhendhi, G. M. Shafiullah, S. Ganesan, A. K. Jeysree, et al., "Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?" Results in Physics, vol. 21, pp. 103817, 2021.
[32] M. Masum, M. A. Masud, M. I. Adnan, H. Shahriar, and S. Kim, "Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for COVID-19 forecasting and management," Socio-Economic Planning Sciences, vol. 80, pp. 101249, 2022.
[33] R. Chandra, A. Jain, and D. S. Chauhan, "Deep learning via LSTM models for COVID-19 infection forecasting in India," PLoS One, vol. 17, no. 1, pp. e0262708, 2022.
[34] A. H. Elsheikh, A. I. Saba, M. A. Elaziz, S. Lu, S. Shanmugan, T. Muthuramalingam, et al., "Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia," Process Safety and Environmental Protection, vol. 149, pp. 223–233, 2021.
[35] S. Shastri, K. Singh, S. Kumar, P. Kour, and V. Mansotra, "Time series forecasting of COVID-19 using deep learning models: India-USA comparative case study," Chaos, Solitons & Fractals, vol. 140, pp. 110227, 2020.
[36] M. B. Braga, R. S. Fernandes, G. N. Souza Jr., J. E. Rocha Jr., C. J. F. Dolácio, and I. S. Tavares Jr., et al., "Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon," PLoS One, vol. 16, no. 3, pp. e0248161, 2021.
[37] F. Cribari-Neto, "A beta regression analysis of COVID-19 mortality in Brazil," Infectious Disease Modelling, vol. 8, no. 2, pp. 309–317, 2023.
[38] R. G. da Silva, M. H. D. M. Ribeiro, V. C. Mariani, and L. S. Coelho, "Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables," Chaos, Solitons & Fractals, vol. 139, pp. 110027, 2020.
[39] M. Marzouk, N. Elshaboury, A. Abdel-Latif, and S. Azab, "Deep learning model for forecasting COVID-19 outbreak in Egypt," Process Safety and Environmental Protection, vol. 153, pp. 363–375, 2021.
[40] M. O. Alassafi, M. Jarrah, and R. Alotaibi, "Time series predicting of COVID-19 based on deep learning," Neurocomputing, vol. 468, pp. 335–344, 2022.
[41] İ. Kırbaş, A. Sözen, A. D. Tuncer, and F. Ş. Kazancıoğlu, "Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN, and LSTM approaches," Chaos, Solitons & Fractals, vol. 138, pp. 110015, 2020.
[42] N. Ayoobi, D. Sharifrazi, R. Alizadehsani, A. Shoeibi, J. M. Gorriz, H. Moosaei, et al., "Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods," Results in Physics, vol. 27, pp. 104495, 2021.
[43] S. Sah, A. Kamerkar, B. Surendiran, and R. Dhanalakshmi, "Predicting the trends of COVID-19 cases using LSTM, GRU, and RNN in India," in Intelligent Data Engineering and Analytics, Singapore: Springer Nature Singapore, 2022.
[44] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting COVID-19 time-series data: A comparative study," Chaos, Solitons & Fractals, vol. 140, pp. 110121, 2020.
[45] E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, et al., "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States," Proceedings of the National Academy of Sciences, vol. 119, no. 15, pp. e2113561119, 2022.
[46] H. M. Scobie, A. G. Johnson, A. B. Suthar, R. Severson, N. B. Alden, S. Balter, et al., "Monitoring incidence of COVID-19 cases, hospitalizations, and deaths, by vaccination status—13 U.S. jurisdictions, April 4–July 17, 2021," MMWR Morbidity and Mortality Weekly Report, vol. 70, no. 37, pp. 1284–1290, 2021.
[47] K. Sreehari, M. Adham, T. D. Cheriya, and R. Sheik, "A comparative study between univariate and multivariate time series models for COVID-19 forecasting," in 2021 International Conference on Computational Performance Evaluation (ComPE), pp. 2021.
[48] M. Sadarangani, B. Abu Raya, J. M. Conway, S. A. Iyaniwura, R. C. Falcao, C. Colijn, et al., "Importance of COVID-19 vaccine efficacy in older age groups," Vaccine, vol. 39, no. 15, pp. 2020–2023, 2021.
[49] K. R. Bhimala, G. K. Patra, R. Mopuri, and S. R. Mutheneni, "Prediction of COVID-19 cases using the weather-integrated deep learning approach for India," Transboundary and Emerging Diseases, vol. 69, no. 3, pp. 1349–1363, 2022.
[50] F. Kamalov, K. Rajab, A. K. Cherukuri, A. Elnagar, and M. Safaraliev, "Deep learning for COVID-19 forecasting: State-of-the-art review," Neurocomputing, vol. 511, pp. 142–154, 2022.
[51] L. Xu, R. Magar, and A. Barati Farimani, "Forecasting COVID-19 new cases using deep learning methods," Computers in Biology and Medicine, vol. 144, pp. 105342, 2022.
[52] "Calcutta: not 'the city of joy'," Gaia: Environmental Information System. Archived from the original on April 27, 2006. Accessed April 26, 2006.
[53] "Weather base entry for Kolkata," Canty and Associates LLC. Archived from the original on September 7, 2011. Accessed April 26, 2006.
[54] A. Dembosky, "Data prescription for better healthcare," Financial Times, p. 19, December 12, 2012. Available: http://www.ft.com/intl/cms/s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK.
[55] B. Feldman, E. M. Martin, and T. Skotnes, "Big data in healthcare: Hype and hope," Dr. Bonnie 360, October 2012. Available: http://www.west-info.eu/files/big-data-in-healthcare.pdf.
[56] M. Fernandes, M. O'Connor, and V. Weaver, "Big data, bigger outcomes," Journal of AHIMA, vol. 83, pp. 38–42, 2012.
[57] M. Kendall and A. Stuart, "The advanced theory of statistics," vol. 3, Griffin, pp. 410–414, 1983.
[58] K. S. Ganguly, S. Modak, A. K. Chattopadhyay, K. S. Ganguly, T. K. Mukherjee, A. Dutta, and D. Biswas, "Forecasting based on a SARIMA model of urban malaria for Kolkata," American Journal of Epidemiology and Infectious Disease, vol. 4, no. 2, pp. 22–33, 2016.
[59] K. S. Ganguly, S. Modak, K. S. Ganguly, and A. K. Chattopadhyay, "Study on temporal effects of urban malaria incidences," International Journal of Statistics in Medical Research, vol. 5, pp. 120–132, 2016.

[60] K. Aho, D. Derryberry, and T. Peterson, "Model selection for ecologists: The worldviews of AIC and BIC," Ecology, vol. 95, pp. 631–636, 2014.
[61] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," Sociological Methods & Research, vol. 33, pp. 261–304, 2004.
[62] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," Giornale dell'Istituto Italiano degli Attuari, vol. 4, pp. 83–91, 1933.
[63] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," Annals of Mathematical Statistics, vol. 19, pp. 279–281, 1948.
[64] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," Biometrika, vol. 65, no. 2, pp. 297–303, 1978.

Authors' Profiles

Krishnendra Sankar Ganguly holds Bachelor's degree in Science in Computer Science from Sam Higginbottom Institute of Agriculture Technology & Sciences (SHIATS), U.P., prior to completion of his three Master's degrees (i) in Computer Science from SHIATS, U.P., (ii) in Computer Application from National Institute of Electronics & Information Technology, Govt of India and (iii) in Technology in Computer Science & Engineering from Birla Institute of Technology, Mesra. He later specialized his knowledge in Data Science, Machine Learning & Deep Learning, Artificial Intelligence & Neural Networks, Predictive Analysis & Forecasting, etc., particularly in Inter-disciplinary Research in Bio-Medical arena. For the last few years, he has been rendering his services in the post of Senior Software Engineer in globally-renowned Enterprises, viz, WIPRO and Ernst & Young GDS.

Dr. Krishna Sankar Ganguly completed his Bachelor's degree in Medical Science (MBBS) from North Bengal University, prior to completion of his two Master's degrees (i) in Management (MBA) from Indira Gandhi National Open University (Govt. of India) and (ii) in Public Health (MPH) from Annamalai University (Tamil Nadu). He later expanded his knowledge through seventeen (17) Diplomas and Certificate Program in Management (PGDHRM & PGDIM), Public Health (DIH & DPH), Statistics (OYPC & BSM) and in several medical disciplines (Sports-Medicine, Rheumatology, Anatomy, Thyroid Disorders, COPD & Asthma, Maternity & Child Health, Optometry, CVD & Stroke, Diabetes, Cancer, Healthcare Quality, etc.) from multiple Universities/ Institutions. He attained a 34 years successful professional career in The Kolkata Municipal Corporation under State Government reaching the upper administrative echelon as Executive Health Officer of Health Department. Simultaneously, he has a proven research career in different capacities (including Visiting Scientist) in Indian Statistical Institute, National Institute for the Orthopedically Handicapped, Jadavpur University, etc., in several scientific arenas of Inter-disciplinary Research, Bio-Engineering, Epidemiology, Malaria, Dengue, PEMF-Therapy, Forecasting, etc. having several original publications.

Dr. Ambar Dutta did his B.Sc. (Honors) in Mathematics from Presidency College, Kolkata and MCA and Ph.D. from Jadavpur University, Kolkata. After serving in the Department of Computer Science and Engineering, Birla Institute of Technology, Mesra for 15 years, he currently works as an Associate Professor at Amity Institute of Information Technology, Amity University, Kolkata. Dr. Dutta authored a book and has published more than 50 papers in reputed national/international journals/conferences. His research interests include Image and Video Processing, Data Analytics, Machine Learning, Information Retrieval, Network Security. He is an active reviewer of many reputed journals like Pattern Recognition Letters (Elsevier), Multimedia Tools and Applications (Springer), IET Image Processing etc. He is a senior life member of various professional bodies

How to cite this paper: Krishnendra Sankar Ganguly, Krishna Sankar Ganguly, Ambar Dutta, "A Comparative Study of Statistical (SARIMA) Vis-À-Vis Some Traditional Machine-Learning and Deep-Learning Techniques to Forecast Malaria Incidences in Kolkata of India", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.68-83, 2025. DOI:10.5815/ijitcs.2025.05.06