Determining Dengue Outbreak Using Predictive Models

Darshan V. Medhane[1](✉) and Varun Agarwal[2](✉)

[1] MVPS's KBT College of Engineering, Nashik, India
darshan.medhane@gmail.com
[2] MIT Academy of Engineering, Pune, India
varunagarwal020@gmail.com

Abstract. Early prediction of contagious and infectious diseases can help health organizations in planning strategies to prevent disease transmission and thus can forestall the outbreak. Several works are there to predict the disease outbreak using climate data but our approach provides better results using univariate model. Our approach is to split the data in terms of variability and volume to find out the best forecasting model for predicting dengue cases with low variability and high-volume data points. For this we have also analyzed the correlation of climate factors with the number of the dengue cases and after comparing the competencies of different forecasting models, we found out that ARIMA is the best suitable model for low variability and high-volume data points with 5.4 RMSE and 3.6 MAE value. The predictive power of these models will be useful to authorities in taking preventive steps.

Keywords: Outbreak · Correlation matrix · Forecasting model · ARIMA

1 Introduction

Coronavirus comes out of nowhere and prompts such countless passing's, assuming we had earlier predictions about its outbreak, such a lot of misfortune doesn't have happened, we could have come with proper preventive measures and saved the lives of many. Coronavirus is one of the recent virus outbreaks but we have been suffering from disease outbreaks from an earlier stage like Spanish influenza killed 40–50 million in 1918, Asian influenza killed 2 million people in 1957, Hong Kong influenza killed 1 million people in 1968, chikungunya, malaria, zika virus and many more. These viruses wherever spread causes obliteration in terms of the lives and economy of the country. These outbreaks have one common thread that is factors contributing to the outbreak. Weather change or small climatic variations comes out to be the key factor for any outbreak of infectious disease transmission [1]. Thus, if prior information for an outbreak is available then it will be easier for doctors to treat patients and the government to make earlier moves.

Dengue fever transmitted by the Aedes mosquitos is greatly influenced by climatic variations throughout the region [2]. Most cases of dengue are not registered because

M. Singh et al. (Eds.): ICACDS 2022, CCIS 1614, pp. 210–221, 2022.
https://doi.org/10.1007/978-3-031-12641-3_18

of asymptotic symptoms revolving around the closely related serotypes namely DEN-1, DEN-2, DEN-3, DEN-4 [3]. This vector-borne disease spreading across the globe and taking life is predominantly dependent on temperature change, as stated by NOAA - 2020 was globally the earth's warmest year and this disease is very sensitive to temperature change. To overcome this problem there is a need for dengue outbreak prediction model which helps in preventing epidemics.

In this paper, we have classified the data in terms of variability and volume and compared the competencies of different forecasting models to find out the best suitable model. For this we have used simple moving average, exponential smoothing, ARIMA, Fbprophet and XGBoost. Rest of the paper is structured as follows: Sect. 2 consists of the previous work in forecasting dengue outbreaks. In Sect. 3, methodology is discussed, Sect. 4 and Sect. 5 discusses algorithms applied and the results for the forecasting model. And in Sect. 6 we have concluded our work with its future aspects.

2 Literature Review

An increase in the number of cases of dengue in several tropical and non-tropical regions gained the interest of many researchers in analyzing the past data and making use of machine learning to forecast the cases. There are several methodologies proposed for the early prediction of dengue outbreaks. Nan et al. [2] proposed a methodology in which they had found five climate conditions responsible for dengue transmission and made predictions using various machine learning models, and found the best accuracy with the XGBoost model. In [4] Mishra et al. proposed a technique in which they used several machine learning models such as neural network, XGBoost, Linear Regression and they founded the best accuracy with Twofolds linear regression. In the above methods, there was no use of time series forecasting models and analyzing patterns in time series is of utmost importance for forecasting. Anggraeni et al. [5] used google trends for increasing the accuracy of prediction by 3% using the ARIMAX model by trying different combinations of p, d, q values. Sillabutra et al. [6] proposed a technique for forecasting dengue morbidity rates using the ARIMA model on finding optimal parameters $p = 3$, $q = 0$, $d = 1$. Anitha et al. in [7] used a decision tree for classification on an individualistic basis whether a person has dengue fever or not, no climate factors were considered in this, only individualistic features were considered. Nakvisut et al. [8] collected open-source data from governmental organizations of Thailand. They had included various climate factors such as average wind speed, maximum temperature, minimum humidity for building a two-step prediction model. For the first step time series forecasting models were used and for the next step used supervised learning, models were there such as linear regression, support vector machine, and neural networks and achieved a root mean squared error of 14.8. Anggraeni et al. [9] used weather data for predicting the number of dengue cases using an artificial neural network (ANN), trying with different combinations of parameters such as learning rate, units of hidden layers, training cycle, and then used Google API for visualization. Manivannan et al. [10] in their paper used a K-means clustering algorithm to make clusters of dengue using dengue serotypes based on the age group. Dinayadura et al. [11] finds the influence of climatic factors on dengue cases and developed a support vector regression model for risk area identification. Jain et al. [12] proposed scenarios where they mentioned the involvement of public traveling in the transmission of dengue cases and mentioned to need to add parameters involving mobility and their transmission. In [13] Makkar et al. developed an early warning system for predicting the number of dengue cases and prominent factors responsible for dengue transmission were pressure and rainfall. Rahmawati et al. [14] uses linear optimization and C-Support vector optimization to predict the dengue fever cases taking different climatic features in account and they had used grid search method for optimizing the parameters and model for better accuracy. Kristianto et al. [15] proposed a technique of combining genetic algorithm and triple exponential smoothing to predict the dengue cases and concluded how GA-TES helped them in increasing accuracy to 8% compared to simple triple exponential smoothing model. Anggraeni et al. [16] used the data of Malang regency and divided it into 3 parts of lowlands, middle and highlands, and found that cases in lowlands and middle lands are higher than highlands cases. They also found that rainfall is also affecting the number of dengue cases because in these villages' dengue cases are higher in the rainy season. Rachata et al. [17] converted the daily data into weekly data and applied entropy technique for feature extraction and to give input to the neural network and use neural network for prediction whether outbreaks occur or not. Chovatiya et al. [18] proposed a technique for the prediction of dengue cases using weather information and used a recurrent neural network that gives an accuracy of 94% as stated by them. Sasongko et al. [19] use various algorithms of backpropagation to predict the early detection of dengue. It included gradient descent, FGS Quasi-Newton, Conjugate Gradient Descent - Powel, Resilient Backpropagation (RB), and Levenberg Marquardt out of which Levenberg Marquardt was the most efficient. Rahim et al. [20] proposed a technology stating using a nonlinear autoregressive moving average with exogenous input and the selection criteria for the parameter of this time series model were AIC, FPE, and Lipschitz, in which they got the accuracy of 88.40%.

3 Methodology

Forecasting algorithm finds pattern in the historical data points and helps in predicting the number of dengue cases in the future time period. The flow of the framework in shown in Fig. 1. We have taken the historical data having weekly number of dengue cases along with the weather conditions in those subsequent areas. We have split the data into four quadrants based on its coefficient of variability vs. volume and targeted the months with more less variability and more volume for data prediction and then performed descriptive analysis on data to have a sanity check on the data and find out the stationarity of the data so that the forecasting algorithms can find patterns in the data while predicting the number of cases. Descriptive analysis of data involves checking for null values, finding modality, skewness and kurtosis of the data to find the outliers in the data and handle them. Also, stationarity of data is checked using the Dicky fuller test to find out if the data is stationary or not using calculation of p-value. Further which we have analyzed different forecasting algorithms underlying their advantages and disadvantages in order to use those models and compared the results of these models to find out the best fit model.

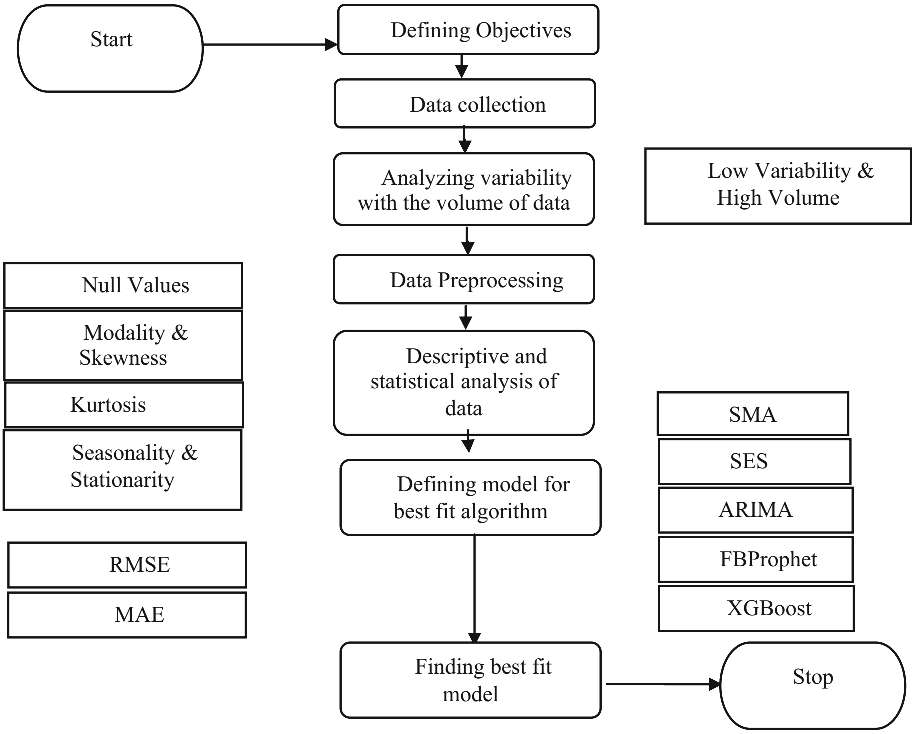

Start
Defining Objectives
Data collection
Analyzing variability with the volume of data
Low Variability & High Volume
Data Preprocessing
Null Values
Modality & Skewness
Kurtosis
Seasonality & Stationarity
Descriptive and statistical analysis of data
Defining model for best fit algorithm
SMA
SES
ARIMA
FBProphet
XGBoost
RMSE
MAE
Finding best fit model
Stop

Fig. 1. Framework for prediction of number of dengue cases

3.1 Dataset

The dataset used for this research purpose consists of data of two cities one is the capital of Puerto Rico i.e. San Juan and another city is Iquitos, a city in Peru.

Data consists of many climate factors such as precipitation, humidity, temperature, relative humidity, average temperature, minimum, and maximum temperature, etc. Data is very little in terms of data points so we have to apply different algorithms and analyze patterns [21].

3.2 Descriptive Analysis and Preprocessing

After data collection process we have done the following descriptive analysis of data to understand the statistics of data and do the preprocessing so that algorithm can be defined based upon the statistical results:

Null Values: Data is checked for any type of null values and we have null values, as our data is less so we cannot simply drop that data. Therefore 3 methods are used in filling the null values that are replacing it with mean, median, or mode, and the corresponding accuracy is checked and we found that median is a better method because there are outliers in the data and mean is very much prone to outliers.

Modality and Skewness: We have unimodal data because we have only one peak in our data and it is positively skewed i.e. skewed towards the right and mean > median. The graph contains unimodality along with the profusion of outliers that is making it rightly skewed.

Kurtosis: It tells us about the outliers and checks the tail whether it is light weighted or heavy weighted.

The kurtosis value for our data is **0.30089** and from this, we can interpret that it is a leptokurtic curve which means it is having a heavy tail and profusion of outliers.

Seasonality: For seasonality, we check if the data is showing any pattern for the summer term or winter term like the cases are high in summer and less in winter or vice versa. That helps in the prediction of future cases.

Stationarity: Models like ARMA, ARIMA, SARIMAX needs the data to be stationery and data is considered to be stationary if it does not show any type of trend or seasonality, we can find this by using Dicky-Fuller Test, and by this, we found that our data is stationary, so there is no need of doing differencing.

Test for Stationarity – Dicky Fuller Test

In this, we calculate the p-value and it has to be less than 0.05 or 5%. If it is greater than 0.05 or 5%, we conclude that the time series has the unit root and accept the null hypothesis.

3.3 Exploratory Data Analysis

EDA for data is necessary as it tells us a story about the data which is useful in selecting what type of processing is required by the data and to understanding the data for choosing optimal algorithm. In our case we have data for two cities i.e. San Juan and Iquitos so

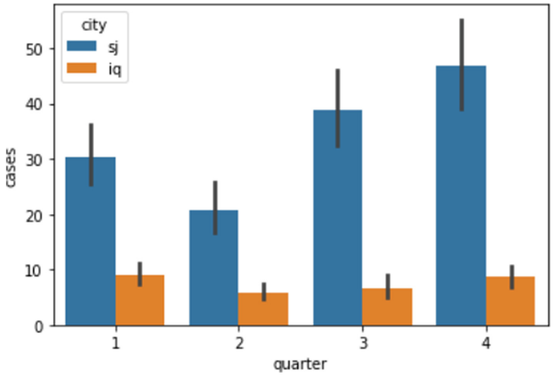


Fig. 2. Dataset consist of two cities so segmented the data and analyzed the number of dengue cases in quarterly manner to see in which quarter there are more number of dengue cases and what is the pattern in both the cities. This also gives us the significant months in which dengue cases were reported more

our data requires segmentation because weather conditions and the number of cases of both these cities are different, so to make sure the data of one city is not affecting the data of another city we will do segmentation (Fig. 1).

San Juan is affected most in quarter 4 i.e. for October, November, December, and Iquitos is most affected at the near end of quarter 4 and quarter 3 i.e. December, January, February, and March (Fig. 2).

We had plotted a scatter plot for finding the relation between temperature and cases in this we can see that as the number of cases increases the temperature also increases so this is showing a correlation and we can use this relation in our multivariate forecasting.

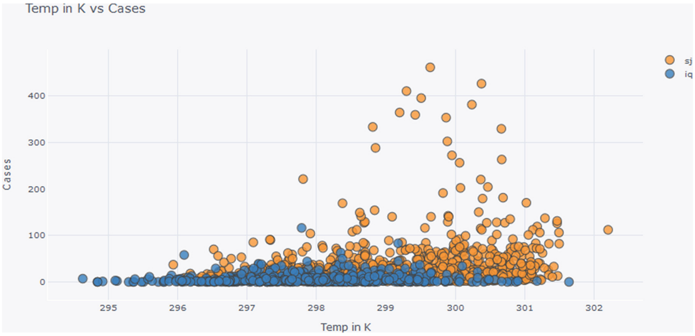


Fig. 3. Relation between temperature and the number of cases shows the positive correlation in data points i.e. in San Juan specifically when the temperature increases number of cases also increases but in a erratic nature

Likewise in the Fig. 4, we can visualize from this graph that cases decrease significantly as the precipitation increase so this is having a negative correlation with the cases. Humidity is having a positive relationship with the number of cases as its relation with the temperature and so these three are climatic factors that we have observed showing correlation with the number of cases.

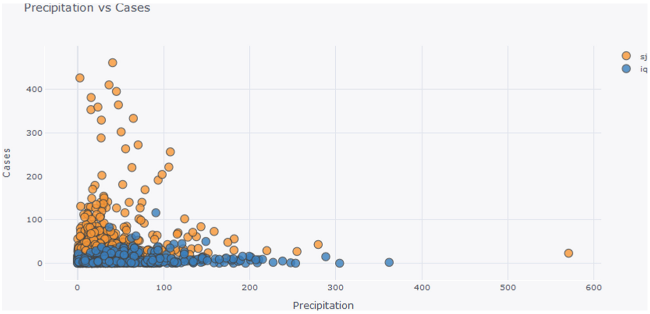


Fig. 4. Relation between Precipitation and the number of cases shows negative correlation

4 Algorithms

For predicting the dengue outbreaks, we have checked the accuracy for following forecasting models to find out the best suitable model discussing their underlying advantages and disadvantages in predicting on historical data.

4.1 Simple Moving Average

SMA (Simple Moving Average) - It is used in forecasting time series data by adding up earlier cases and dividing them by the total number of cases. If there is an unusual change in the cases it is preferred as it shows a true average over time. The disadvantage of SMA is it fails to accurately reflect the most recent trends and does not consider any other features.

4.2 Simple Exponential Smoothing

SES (Simple Exponential Smoothing) – It is used in forecasting time series data that does not have any trend or seasonality. The advantage of SES is it requires less data storage because in this we work on only two factors unlike SMA, also it is very simple to implement and powerful because of its weighting process. The disadvantage of this process is it lags and is non-adaptive i.e. it does not consider dynamic changes.

$$S_t = \alpha y_{t-1} + (1 - \alpha)S_{t-1} \tag{1}$$

4.3 Auto-regressive Integrated Moving Average

It is used in the prediction of time series data. It comprises 3 main terms i.e., Auto-Regressive which is lags of variables itself, integrated which is several steps required to make data stationery, and Moving Average Lags which are lags of previous information. The assumption is taken by ARIMA that the data is stationary.

The advantage of ARIMA is they are more accurate and reliable in case of erratic data.

The disadvantage is it captures only linear relationships; hence, a neural network model or genetic model could be used if a nonlinear association (ex: quadratic relation) is found in the variables.

$$y' t = c + \phi(1) y' t - 1 + \cdots + \theta(1)\varepsilon(t-1) + \cdots + \theta(q)\varepsilon(t-q) + \varepsilon(t) \tag{2}$$

To find the optimal parameters for ARIMA model ACF (Auto correlation) and PACF (partial auto-correlation plots are used).

An autocorrelation plot is a plot of total correlation between different lag functions. If there is a positive autocorrelation at lag 1 then we use the autoregressive model. If there is a negative autocorrelation at lag 1 then we use the moving average model.

PACF is the correlation between two variables under the assumption that we know and take into account the values of some other sets of variables. If this model drops off at lag n, then use an AR(n) model and if the drop in PACF is more gradual then we use the moving average term.

4.4 FbProphet

Fbprophet is a time series forecasting algorithm developed by Facebook. It takes four factors into account i.e. seasonal, trends, holidays, and error or event effect. In this model non-linear trends are fit with yearly, weekly, seasonality, and holiday effects

$$\mathrm{Y(t) = g(t) + s(t) + h(t) + \varepsilon t} \tag{3}$$

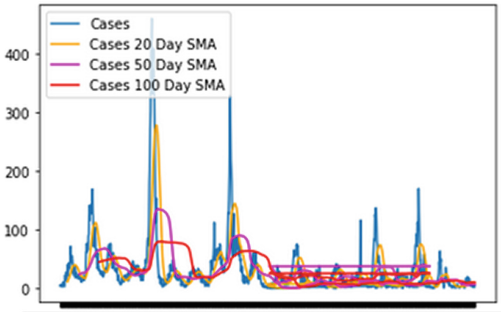
The Fbprophet model is effective in cases when data is having seasonality's, outliers in form of important events or Holiday which had led to a high number of cases, and historical trend changes and data must be at least of 1 year.

Advantages – It handles data with uneven time intervals, handles the null values, handles seasonality, and works well by default setting.

Disadvantages – Fails in forecasting erratic data, cannot handle exogenous variables, multiplicative models and data in prophet need to be fees in a pre-defined format.

5 Algorithms and Results

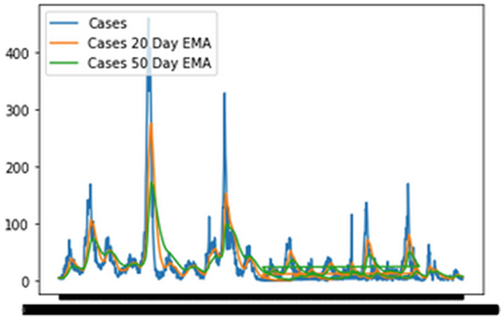
In the above graph, we had taken the training data frame and plotted the earlier cases, moving average on 20, 50, 100 rolling windows and we can see that 20-day rolling window is performing better in forecasting and it can catch patterns but for rolling window 50 and 100 we are seeing some erratic patterns that are worsening the forecast (Fig. 5).



[redacted]

Fig. 5. Simple moving average model shows that there is significant error in the number of cases and the 20-, 50- and 100-days moving average

In exponential smoothing, we can see the 20 and 50-day exponential moving average and in this, we have tried with different values of the alpha parameter to reduce the error in forecasting and this also 20-day rolling window give us good results but not better (Fig. 6).



[redacted]

Fig. 6. Exponential smoothing model results shows how there is difference in the actual number of cases and the cases forecasted through exponential smoothing i.e. 20 days exponential moving average and 50 days moving average

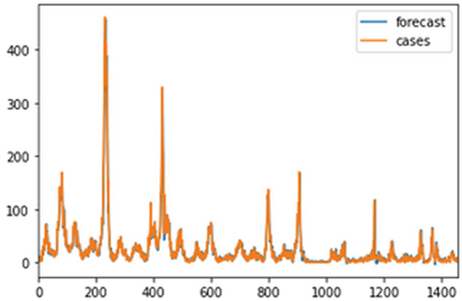


Fig. 7. Forecasted values are mostly overlapping with the number of dengue cases which is stating how accurately ARIMA has found patterns in the data

ARIMA takes three parameters p, d, q in which p and q are determined with ACF and PACF plots and we have to choose the combination of p, d, q which have minimum AIC (Acyle information criteria), and for that we have to try a different combination of p, d, q values or use auto Arima model that takes the range of p and q values and searches for best parameters, one advantage for using auto-Arima model is it removes seasonality, stationarity from the data after applying the seasonality and stationarity check (Table 1) (Fig. 7).

Simple moving average having 22.7 RMSE and 16.3 MAE and exponential smoothing having 14.01 RMSE and 7.02 MAE showing very high error because they are not able to analyse the dengue cases patterns in the past year in specific months. Fbprophet showing high inaccuracy because of the erratic nature of the data points and not able to predict the data points. But we can see that ARIMA is performing better than other time series forecasting models and Ensemble models i.e. XGBoost because in ARIMA it is taking the correlation of months of this year with months of the past year. The values of p, d, q are chosen by looking at the partial autocorrelation plot, autocorrelation plot, such that final model is having minimum AIC value and used in almost correct prediction of the number of dengue cases.

Table 1. Results from algorithms used for forecasting

S. no.	Algorithms used	RMSE	MAE
1	Simple Moving Average	22.68	16.29
2	Exponential Smoothing	14.01	7.024
3	ARIMA	5.38	3.59
4	FBProphet	18.58	13.1
5	XGboost	17.71	12.83

6 Conclusion

Analysis of the data showed how climate factors are affecting the number of cases and maintain a direct and indirect relationship. From our approach we found that data which lie in less variability and more volume quadrant is of the months for which dengue cases are highly reported and have a strong correlation with the climate factors but after applying the aforementioned algorithms, we found out that ARIMA is the best fit model for this type of data with 5.4 RMSE and 3.6 MAE. There is scope of reducing this error if the data points were more so that we can find out the missing patterns in the data. Currently, we had prepared the model that contains only two cities, so it requires segmentation at one level only and then runs the best fit model, which will give the algorithm the most optimal results. For future scope, we can expand our model to the data including multiple granularities, and segment the data area wise to be optimally utilized by the concerned authorities.

References

1. Andrick, B., Clark, B., Nygaard, K., Logar, A., Penaloza, M., Welch, R.: Infectious disease and climate change: detecting contributing factors and predicting future outbreaks. In: 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings, IGARSS 1997. Remote Sensing - A Scientific Vision for Sustainable Development, Singapore, vol. 4, pp. 1947–1949 (1997). https://doi.org/10.1109/IGARSS.1997.609159
2. Nan, J., et al.: Using climate factors to predict the outbreak of dengue fever. In: 2018 7th International Conference on Digital Home (ICDH), Guilin, China, pp. 213–218 (2018). https://doi.org/10.1109/ICDH.2018.00045
3. Ebi, K.L., Nealon, J.: Dengue in a changing climate. Environ. Res. **151**, 115–123 (2016). ISSN 0013-9351
4. Mishra, V.K., Tiwari, N., Ajaymon, S.L.: Dengue disease spread prediction using twofold linear regression. In: 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, pp. 182–187 (2019). https://doi.org/10.1109/IACC48062.2019.8971567
5. Anggraeni, W., Aristiani, L.: Using Google Trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia. In: 2016 International Conference on Information & Communication Technology and Systems (ICTS), Surabaya, pp. 114–118 (2016). https://doi.org/10.1109/ICTS.2016.7910283

6. Sillabutra, J., Soontornpipit, P., Viwatwongkasem, C., Satitvipawee, P., Phuthomdee, S.: Forecasting model for dengue morbidity rate in Thailand. In: 2018 International Electrical Engineering Congress (iEECON), Krabi, Thailand, pp. 1-4 (2018). https://doi.org/10.1109/IEECON.2018.8712202
7. Anitha, A., Wise, D.C.J.W.: Forecasting dengue fever using classification techniques in data mining. In: 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 398–401 (2018). https://doi.org/10.1109/ICSSIT.2018.8748864
8. Nakvisut, A., Phienthrakul, T.: Two-step prediction technique for dengue outbreak in Thailand. In: 2018 International Electrical Engineering Congress (iEECON), Krabi, Thailand, pp. 1–4 (2018). https://doi.org/10.1109/IEECON.2018.8712258
9. Anggraeni, W., et al.: Artificial neural network for health data forecasting, case study: number of dengue hemorrhagic fever cases in Malang regency, Indonesia. In: 2018 International Conference on Electrical Engineering and Computer Science (ICECOS), Pangkal Pinang, pp. 207–212 (2018). https://doi.org/10.1109/ICECOS.2018.8605254
10. Manivannan, P., Devi, P.I.: Dengue fever prediction using K-means clustering algorithm. In: 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, pp. 1–5 (2017). https://doi.org/10.1109/ITCOSP.2017.8303126
11. Dinayadura, N.S., Mikler, A.R., Muthukudage, J.: An efficient approach of outbreak preparedness for dengue. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, p. 327 (2017). https://doi.org/10.1109/ICHI.2017.16
12. Jain, R., Sontisirikit, S., Iamsirithaworn, S., et al.: Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC Infect. Dis. **19**, 272 (2019)
13. Makkar, G.: Real-time disease forecasting using climatic factors: supervised analytical methodology. In: 2018 IEEE PuneCon, Pune, India, pp. 1-5 (2018). https://doi.org/10.1109/PUNECON.2018.8745369
14. Rahmawati, D., Huang, Y.: Using C-support vector classification to forecast dengue fever epidemics in Taiwan. In: 2016 International Conference on System Science and Engineering (ICSSE), Puli, pp. 1–4 (2016). https://doi.org/10.1109/ICSSE.2016.7551552
15. Kristianto, R.P., Utami, E.: Optimization the parameter of forecasting algorithm by using the genetical algorithm toward the information systems of geography for predicting the patient of dengue fever in district of Sragen, Indonesia. In: 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, pp. 45–50 (2017). https://doi.org/10.1109/ICITISEE.2017.8285548
16. Anggraeni, W., et al.: Modelling and forecasting the dengue hemorrhagic fever cases number using hybrid fuzzy-ARIMA. In: 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), Kyoto, Japan, pp. 1–8 (2019). https://doi.org/10.1109/SeGAH.2019.8882433
17. Rachata, N., Charoenkwan, P., Yooyativong, T., Chamnongthal, K., Lursinsap, C., Higuchi, K.: Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network. In: 2008 International Symposium on Communications and Information Technologies, Lao, pp. 210–214 (2008). https://doi.org/10.1109/ISCIT.2008.4700184
18. Chovatiya, M., Dhameliya, A., Deokar, J., Gonsalves, J., Mathur, A.: Prediction of dengue using recurrent neural network. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, pp. 926–929 (2019). https://doi.org/10.1109/ICOEI.2019.8862581

19. Sasongko, P.S., Wibawa, H.A., Maulana, F., Bahtiar, N.: Performance comparison of Artificial Neural Network models for dengue fever disease detection. In: 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), Semarang, pp. 183–188 (2017). https://doi.org/10.1109/ICICOS.2017.8276359
20. Abdul Rahim, H., Ibrahim, F., Taib, M.N.: A novel prediction system in dengue fever using NARMAX model. In: 2007 International Conference on Control, Automation and Systems, Seoul, pp. 305–309 (2007). https://doi.org/10.1109/ICCAS.2007.4406927
21. Dengue Forecasting. https://dengueforecasting.noaa.gov/