

Chapter 5

Machine Learning Techniques for Predicting Dengue Outbreak



**M. B. Anuranjan, Chandana Divya Vani, Chanpreet Singh,
Samarjeet Barman, Kuldeep Chaurasia, and P. V. Arun**

1 Introduction

Dengue is an exponentially emerging commonly prone viral disease in many parts of the world. Dengue flourishes in urban poor areas, suburbs and the countryside but also affects more affluent neighbourhoods in tropical and subtropical countries. Apart from causing a severe flu-like illness, sometimes it is prone to cause a lethal complication called severe dengue. According to World Health Organization (WHO), 50–100 million infections are now estimated to occur annually [1]. The *Aedes aegypti* mosquito transmits the viruses that cause dengue. Because dengue is carried by mosquitoes, the transmission of dengue is related to meteorological and environmental data such as temperature, precipitation and vegetation. A vaccine to prevent dengue is licenced and available in some countries. Surprisingly, the vaccine manufacturer has announced in 2017 that people who have not been previously infected and received the vaccine may be at risk of developing severe dengue if they get infected following the vaccination [2]. The backbone for the treatment of dengue is supportive

M. B. Anuranjan (✉)
Rajalakshmi Engineering College, Chennai, Tamil Nadu, India
e-mail: anuranjan.mb.2018.cse@rajalakshmi.edu.in

C. Divya Vani
JNTUK, Kakinada, Andhra Pradesh, India

C. Singh
Amrita University, Kollam, Kerala, India

S. Barman
IIIT Manipur, Heingang, Manipur, India

K. Chaurasia
Bennett University, Greater Noida, India

P. V. Arun
Indian Institute of Information Technology, Sri City, Chittoor, India

care [3]. It becomes vital to recognize the three phases of dengue [4]: febrile, critical and recovery. Hence, serious efforts are required to control and prevent this disease. This makes predictions on dengue outbreaks very important. With the help of this prediction, health departments across the world can take preventive measures to combat dengue fever before the outbreak begins, saving millions of lives. The objective of the research work is to find the best algorithm to predict the spread of Dengue in the area under study.

2 Related Work

Several approaches have been used to predict dengue outbreaks. Rachata et al. [5] worked on the use of artificial neural networks with an entropy method to build a predictive model for dengue outbreaks in Thailand. A study in Sri Lanka, using past patterns of weather and past dengue cases as inputs, introduced an artificial neural network (ANN) to predict the dengue outbreak in the Kandy district [6]. Cheong et al. [7] showed that land-use factors other than human settlements, like different types of agricultural land, water bodies and forest, can be associated with reported dengue cases in the state of Selangor, Malaysia, and used boosted regression to account for nonlinearities and interactions between these factors. Dharmawardana et al. [8] proposed that the mobility of humans has a significant effect on the outset of dengue to the immunological dengue ‘naive’ region. The data set was derived using mobile network big data in Sri Lanka. The predictions were made using ANN and XGBoost. Recent work of Muhilthini et al. [9] proposed a dengue possibility forecasting model; the data set contains information about several dengue cases observed every week for several years in any country. It contains details about the conditions of weather like precipitation amount, temperature, humidity and so on using GBR to find the dependencies in the given training data set and predict the amount of dengue cases for the given week and year of a country in the test data set. Tate et al. [10] proposed a model for the prediction of dengue, diabetes and swine flu using random forest. The main aim of this model is to predict the disease by using the symptoms taken from patients, and to recommend a specialized doctor, from this, the risky cases of that particular disease in a week of that particular area was also calculated. Ong et al. [11] used random forest regression to predict the risk rank of dengue transmission in 1 km grids, with dengue, population, entomological and environment data in Singapore. More than 80% of the observed risk ranks fell within the 80% prediction interval.

3 Study Area and Data Resources

The two cities considered for this experiment are San Juan, Puerto Rico, and Iquitos, Peru. San Juan is the capital city of Puerto Rico and also happens to be the most populous municipality. Iquitos is the capital city of the Mayan Province and Loreto region

and is claimed to be the ninth most populous city of Peru. (Fig. 1). The data used in this study were collected by various U.S Federal Government Agencies (including the Centers for Disease Control and Prevention, the National Oceanic and Atmospheric Administration and the U.S. Department of Commerce). The consolidated data set was acquired from the “Data Download” section of the openAI competition “DengAI: Predicting Disease Spread” hosted by drivendata.org. The data set was a combination of meteorological data including the features. Some of the important features are city abbreviations: sj for San Juan and iq for Iquitos, week_start_date, maximum temperature, minimum temperature, average temperature, total precipitation, mean air temperature, pixel southeast of city centroid, pixel northeast of city centroid etc.

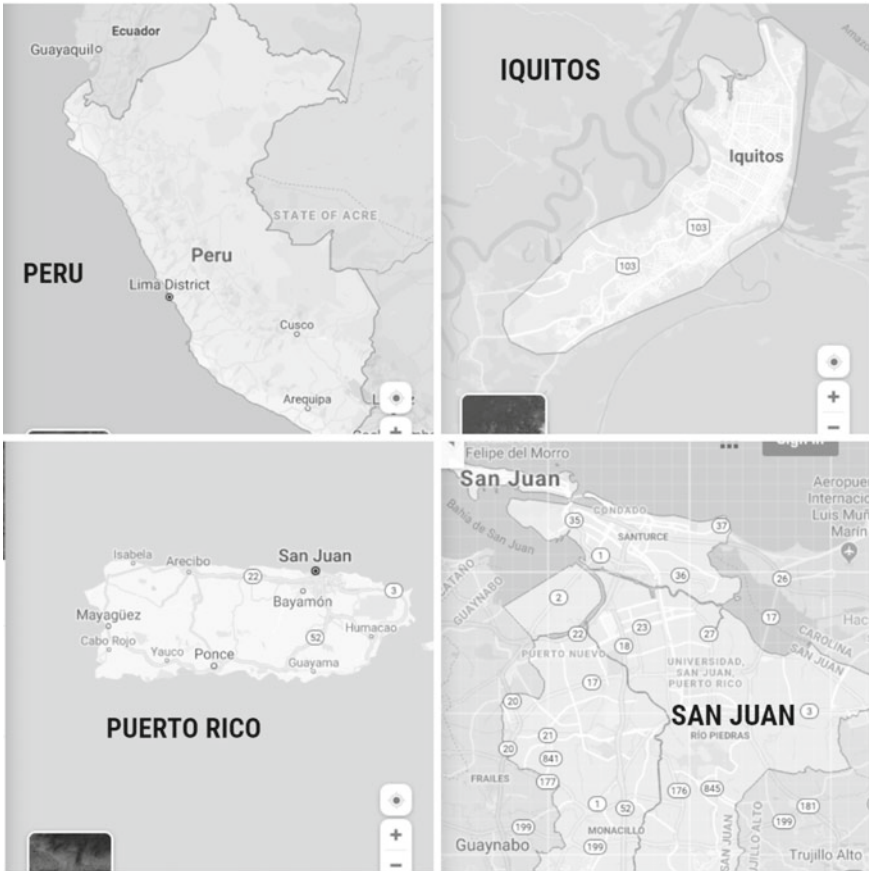


Fig. 1 Study area [13]

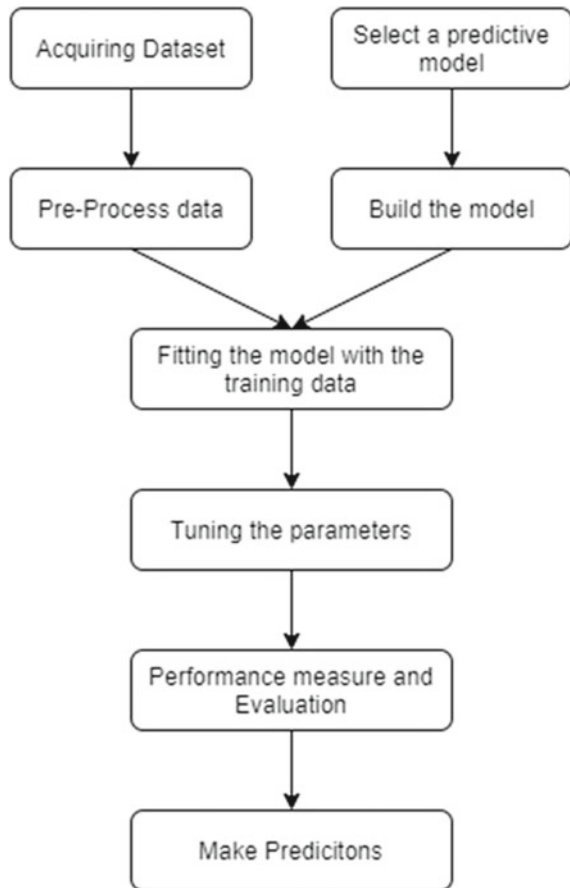
4 Methodology

The methodology involved seven steps, namely acquiring the data set, preprocessing the data, selecting a relevant predictive model, building the model, fitting and training the model with the training data, tuning the parameters for optimal performance and predicting the values (Fig. 2).

4.1 Dataset Preprocessing

The acquired data had to be preprocessed for maximum performance of the models. Various data preprocessing techniques were used in this study. To begin with, the data set had to be divided based on the two cities. A new feature called “month”

Fig. 2 General methodology applied to make various predictive models

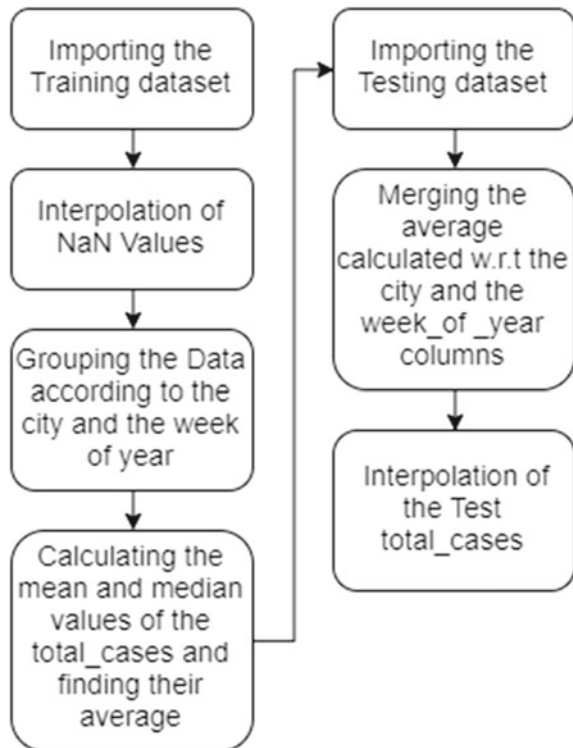


had to be extracted from the “week_start_date” since the latter is a little harder to work with. The Not a Number (NaN) values were replaced with the mean of their respective columns. It can be noticed that a few features associated with temperature were recorded in Kelvin and a few others in Celsius. The features were all converted to Celsius.

4.2 Interpolation

Interpolation is the process of acquiring a function from a set of data points such that the function passes through all the given data points and can be used to appraise data points in-between the given ones. Using interpolation (Fig. 3) and the weekly pattern of the recorded cases, it is possible to make a plausible prediction. To attain a weekly pattern, the data from the training set can be grouped by the week_of_year column. The mean and median of this grouped data are averaged to get an appreciable week pattern. The pattern is then later used to predict the total_cases reported in the testing data. Interpolation of the predicted total_cases is done for accurate results.

Fig. 3 Algorithm used to make the interpolation model



4.3 Time Series Forecasting with Random Forest

Random forest is a subset of the supervised learning algorithms which uses an ensemble learning method. It can be used for both classification and regression. The trees in random forests are made to run in parallel meaning there is not any interaction between these trees. A random forest combines many decision trees. The differences lie in the number of features that can be split at each node and the randomness added by random selection of sample data from the original data by the decision trees to avoid over-fitting. A time series is a vector of values that are indexed by time. Sometimes, it is necessary to perform some preprocessing to make it comprehensible (Fig. 4).

This algorithm involved the prediction of total cases using time series forecasting with random forest. The data set after being imported was preprocessed. The data preprocessing involved replacing the NaN values using the forward and backward fill method.

It was necessary to split the data set based on the cities for accurate predictions. The next step involved introducing time series based lag which served as the base for the time series forecasting of the model. A random forest regressor with the necessary parameters was built. The next step involved the fitting of the data set to the model. Two parameters were tuned for better performance. The two parameters considered are `n_estimators` and `max_depth`. After trying out various values, the above-tabulated values (Table 1) were considered for the cities as using these parameters gave the

Fig. 4 Algorithm used with time series prediction with random forest regression

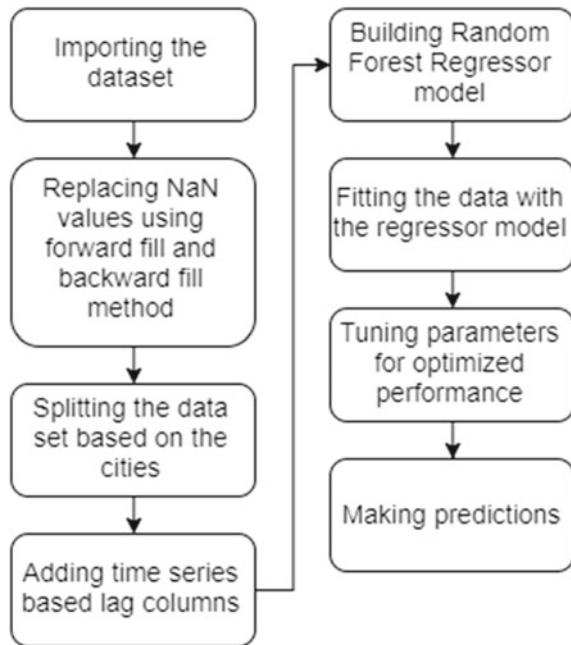


Table 1 Details of various parameters tuned in the random forest model

Parameter tuning			
Parameter	Values	San Juan	Iquitos
n estimator	10,20,50,100,200,300,500	20	100
max depth	10,15,20,40,50,100	10	10

best result following which the model has trained again with the tuned parameters and used to predict the total_cases. The mean absolute error computed (using a time series-based algorithm) for both the cities was better than the MAE calculated using the normal random forest algorithm.

4.4 Gradient Boosting for Time Series Prediction

Gradient boosting is one of the most powerful techniques for building predictive models. It is a machine learning technique widely used for regression and classification problems. Gradient boosting aims at producing an ensemble of weak prediction models. The algorithm repetitively leverages the patterns in residuals and bolster a model with weak prediction to improve it. This algorithm involved the prediction of total_cases using time series forecasting with gradient boosting. The data set after being imported needed to be preprocessed. The data preprocessing involved replacing the NaN values using the forward and backward fill method. It was necessary to split the data set based on the cities for more accurate predictions. The next step involved introducing time series-based lag which served as the base for the time series forecasting of the model. A gradient boosting regressor with the necessary parameters was built (Fig. 5).

The next step involved the fitting of the data set to the model. Four parameters were tuned for better performance (Table 2).

4.5 Performance Measures

The performance of both models has been assessed by computing the mean absolute error (MAE). The mean absolute error (MAE) [12] measures the closeness of forecasts predictions to the actual outcomes. It is given by (Eq. 1):

$$\text{MAE} = \frac{1}{n} \sum_1^n |y_i - y_{\hat{i}}| = \frac{1}{n} \sum_1^n |e_i|. \quad (1)$$

$$\text{AE} = |e_i| = |y_i - y_{\hat{i}}| \quad (2)$$

where actual = y_i and predicted = $y_{\hat{i}}$.

Fig. 5 Algorithm used in implementing the gradient boosting regression model

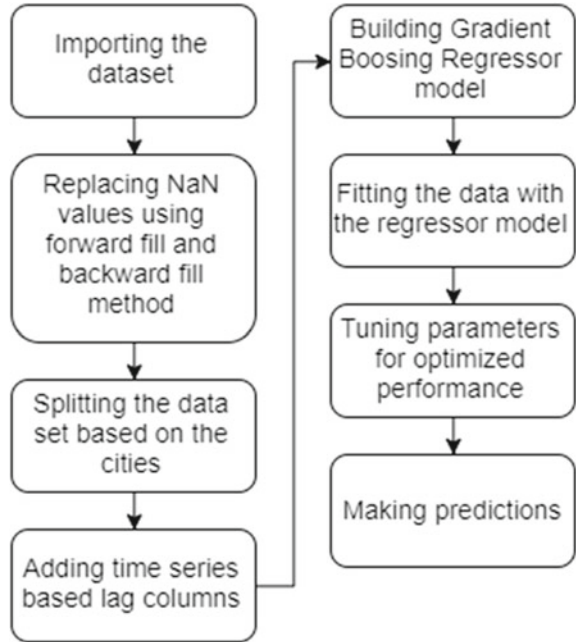


Table 2 Various parameters tuned in the gradient boosting model

Parameter tuning			
Parameter	Values	San Juan	Iquitos
max iter	10,20,50,100,200,300,500	50	10
max leaf nodes	10,16,32,64,128,256	10	16
max depth	4,8,16,32	8	10
max bins	10,16,32,64,128,256	10	32

5 Results and Discussion

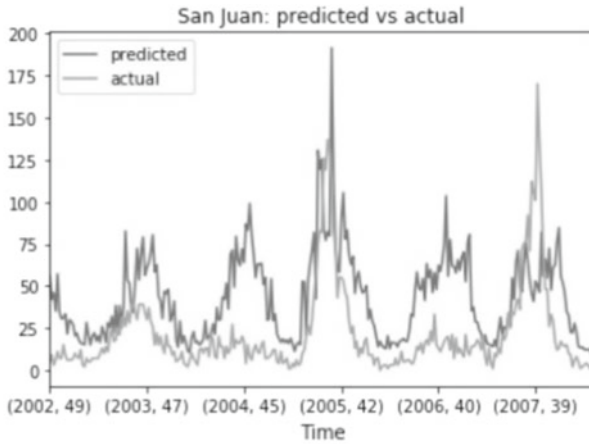
Time series-based prediction is a good alternative for ordinary predictive models. The three algorithms used were manipulation using interpolation, times series-based random forest and gradient boosting. Although these algorithms perform well, the time series-based gradient boosting algorithm seems to have the best of results.

The MAE of this algorithm is better than those of others. The respective MAE values have been tabulated (Table 3).

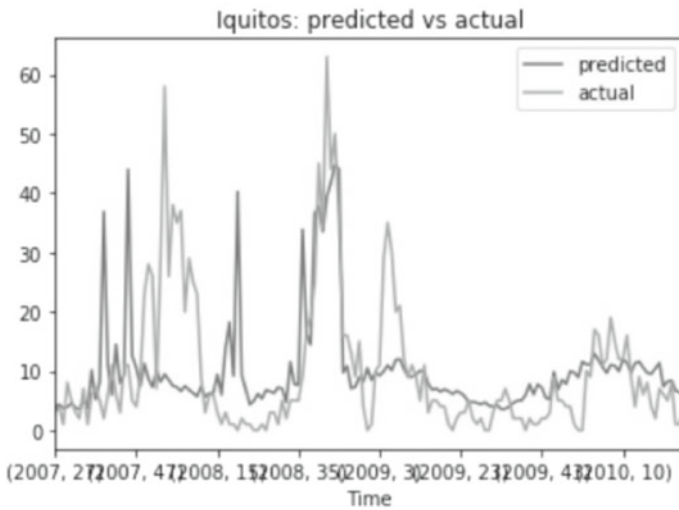
The random forest regressor seems to work the best in this particular problem. It can be observed from the graph (Fig. 6) that the predicted values were almost the same as the reported cases in most of the instances. Although the values might not

Table 3 Comparison of results of various predictive model

Comparison of MAE using different algorithms		
Algorithms	San Juan	Iquitos
Random forest	26.66	6.7
Gradient boosting	24.11	7.36
Interpolation	25.98	25.98



(a)



(b)

Fig. 6 Plotting result using random forest: **a** San Juan region, **b** Iquitos region

be very close, the regressor had done a good job in predicting and understanding the pattern involved.

The gradient boosting algorithm does not seem to have worked the best in this situation (as shown in Fig. 7). The algorithm did a good job of predicting the number of reported cases in Iquitos.

The algorithms might have made better results with a deep analysis of the data set and more fine-tuning of the model.

Overall, the gradient boosting algorithm seems to have worked the best among the other algorithms used. The other algorithm used involves a random forest regressor

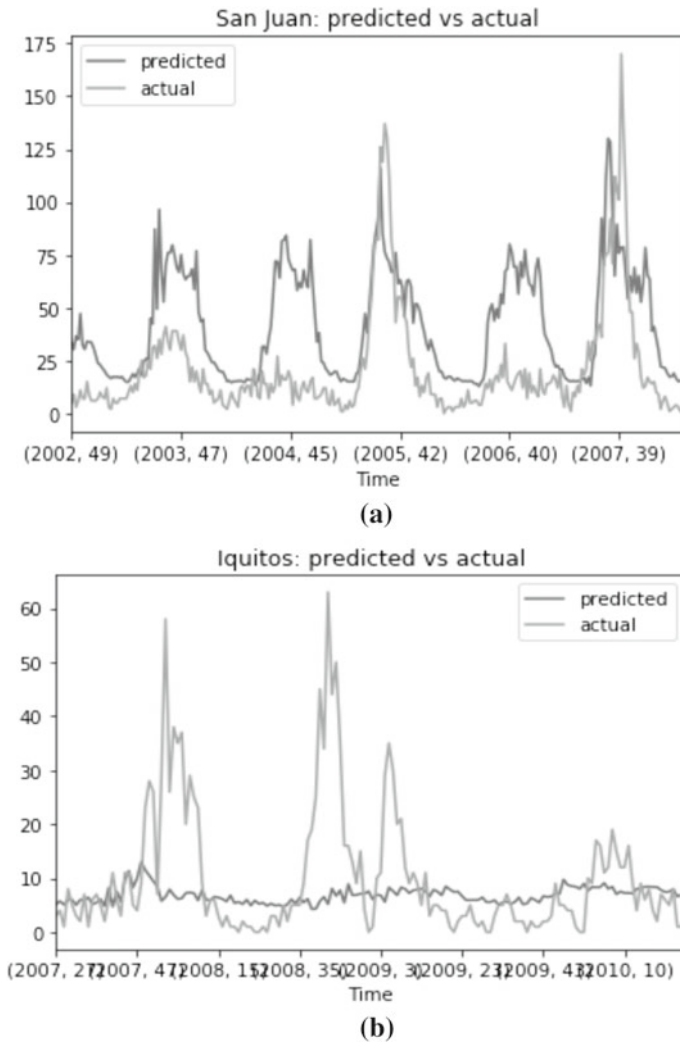


Fig. 7 Plotting result using gradient boosting: **a** San Juan region, **b** Iquitos region

and interpolation technique which seems to be providing an acceptable result. The following limitations made the process of achieving better results harder and also undermined the scope of using deep neural networks to make predictions: data set given data is limited or less, two cities are located in different geographical areas, and dengue occurs at particular months of the year. This result can further be developed by using models that are known to work better for time series-based data like the ARIMA model.

6 Conclusion

This study considered three different modelling techniques, namely interpolation, gradient boosting regression and random forest regression, to predict the amount of dengue cases reported in two cities. Parameters were tuned and adjusted for the most optimal performance. Comparisons of results were made based on mean absolute error (MAE). The performance was analysed, and the result points out that the gradient boosting regression performs significantly better than the other models and is therefore considered to be a better approach. Future results can be improved by obtaining large amounts of meaningful data and implementing better models associated with time series predicting.

Acknowledgements The authors would like to thank Bennett University for supporting the project and providing all necessary facilities. The authors would also like to thank the following U.S Federal Government Agencies for providing the data—Centers for Disease Control and Prevention (CDC), National Oceanic and Atmospheric Administration and the U.S. Department of Commerce. The research problem was proposed through a series of competitions held by drivendata.org. The authors would like to acknowledge drivendata.org for this opportunity.

References

1. Safitri MD, Yusniar H (2019) Association between environmental factors and the presence of mosquito larvae to dengue hemorrhagic fever (DHF) in Karimunbesar Island, Indonesia. *Int J Health, Educ Soc (IJHES)* 2(12):18–25
2. Halstead SB, Deen J (2002) The future of dengue vaccines. *The Lancet* 360(9341):1243–1245
3. Wallace D, Canouet V, Garbes P, Wartel TA (2013). Challenges in the clinical development of a dengue vaccine. *Curr Opin Virol* 3(3):352–356
4. Gubler DJ (1998) Dengue and dengue hemorrhagic fever. *Clin Microbiol Rev* 11(3):480–496
5. Rachata N, Charoenkwan P, Yooyativong T, Chamnongthai K, Lursinsap C, Higuchi K (2008) Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network. *Iscit* 210–214
6. Herath PHMN, Perera AAI, Wijekoon HP (2014) Prediction of dengue outbreaks in Sri Lanka using artificial neural networks. *Int J Comput Appl* 101(15):1–5
7. Cheong YL, Leitão PJ, Tobia L (2014) Assessment of land use factors associated with dengue cases in Malaysia using boosted regression trees. *Spatial Spatio-temporal Epidemiol* 10(2014):75–84

8. Dharmawardana KGS, Lokuge JN, Dassanayake PSB, Sirisena ML, Fernando ML, Perera AS, Lokanathan S (2017) Predictive model for the dengue incidences in Sri Lanka using mobile network big data. In: 2017 IEEE international conference on industrial and information systems (ICIIS), pp 1–6. IEEE
9. Muhilthini P, Meenakshi BS, Lekha SL, Santhanalakshmi ST (2018) Dengue possibility forecasting model using machine learning algorithms. *Int Res J Eng Technol* 5(3):1661–1665
10. Tate A, Gavhane V, Pawar J, Rajpurohit B, Deshmwch GB (2017) Prediction of dengue diabetes and swine flu using a random forest classification algorithm. *Int RJ Eng Tech* 4(2017):685–690
11. Ong J, Xu L, Rajarethinam J, Kok SY, Liang S, Tang CS, Cook AR, Ng LC, Yap G (2018) Mapping dengue risk in Singapore using random forest. *PLoS Neglected Trop Diseases* 12(6):e0006587
12. Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res* 30(1):79–82
13. Google maps. [Online] Available: <https://www.google.com/maps>