Prediction of Outbreak Periods of Dengue in Baguio City, Philippines using Machine Learning Classification Models

Jozelle C. Addawe, Richelle Ann B. Juayong, and Jaime D.L. Caro

Department of Computer Science, University of the Philippines Diliman
Diliman, Quezon City, Philippines
jcaddawe@up.edu.ph

Abstract. Detection of possible disease outbreak is a vital role of disease surveillance. Previous studies on dengue in Baguio City, Philippines include exploratory and spatiotemporal analysis, modeling and forecasting methods, but lacks approaches for detection of outbreak. This study aims to obtain a model that may be used to predict outbreaks using variables that have been shown in literature to affect the increase of dengue cases. Machine learning classifiers such as the random forest, decision trees and gradient boosting methods are tested for their performance in classifying outbreak and non-outbreak periods in five barangays of Baguio City in 2019 to 2020. Results have shown that the random forest classifier outperforms the other two classifiers in terms of prediction accuracy, with at least 75% accuracy for predicting outbreak months. The model is further improved with average cases, relative humidity, temperature and lagged values of dengue as input variables to the random forest classifier.

Keywords: dengue, outbreak, machine learning, classification

1 Introduction

As part of disease surveillance, it is important that possible outbreaks are earlier detected to prevent the progression of the disease and stop it from spreading and increasing morbidity and mortality in neighboring areas. Not only will possible outbreak detection help prevent the increase in disease cases, but also serve as an evaluation or a precautionary measure for identifying underlying conditions that seem to be 'abnormal' compared to historical records of the disease, and making it useful for designing intervention plans.

In this research, we aim to contribute to previous research and to the surveillance of dengue in Baguio City, Philippines by identifying possible outbreak periods of dengue within local communities based on historical records. Recent studies on dengue in Baguio City have performed exploratory data analysis, modeling, and time series forecasting for predicting dengue cases in the entire city. Predictive analytics are limited to basic interpolation methods, linear regression and multiple regression, ARIMA and SARIMA modeling.

J. Caro et al. (eds.), *Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2023)*, Atlantis Highlights in Computer Sciences 20,
https://doi.org/10.2991/978-94-6463-388-7_23

A more recent study in [1] explored the use of machine learning models for the prediction of dengue cases in the City. Nonetheless, in depth application of these predictive methods within the local communities as well as possible outbreak detection of dengue remain unexplored. Moreover, while forecasting may be a useful tool for disease surveillance, its accuracy is only reliable for short term forecasts of the expected number of disease cases. Following the importance of outbreak detection and the lack of outbreak detection research to the region of study, this research evaluates and identifies the best machine learning model to be used in detecting time periods when dengue outbreaks in barangays of Baguio City, Philippines may possibly occur.

The use of machine learning (ML) models have been previously used and have been shown to have potential for the dengue surveillance and prediction of dengue outbreaks [9][3][2][5][4]. Machine learning classification models were tested in a study by Salim, et.al. [9] for their performance in detecting outbreak and outbreak weeks. Similar to their study, this study explores the performance of machine learning classification models to identify outbreak and non-outbreak periods of dengue in 5 barangays of Baguio City with the highest dengue cases yearly.

2 Previous Studies on Dengue Outbreak Detection

Dengue outbreak detection serves as an early warning system in the surveillance of dengue. According to Baharom et. al., in designing such systems, previous case records or alarm indicators such as meteorological, epidemiological, entomological, population and socioeconomic data may be used [2]. Among these, several papers have shown that meteorological data and record of previous cases are the best indicators for dengue occurrence. Common methods used in designing early warning systems including outbreak detection are statistical and machine learning methods.

A previous study on dengue outbreak detection in Baguio City was conducted by Marigmen et. al. in 2021 [7]. In their paper, possible dengue outbreaks from 2016 to 2018 in Baguio City were determined using the reproduction number, R_0, which was computed from the weekly aggregated dengue cases in the city. According to their analysis, notable peaks of the reproduction number exceeding 1 occurred during the end of June to early July of 2016, second week of July 2017 and first week of July in 2018. During these times, the infection rate in the city was at least 4 times higher than the recovery rate. Following their analysis of the observed trends, it was predicted that a dengue outbreak call may occur in the city during late June to early July of 2019. In their paper, an outbreak was defined based on the increase in reproduction number of dengue which was observed to occur during the same time of every year. However, according to the CDC, an outbreak is the occurrence of a disease that is greater than the expected occurrence at a particular time and space. Thus, following this definition, these months may not necessarily be considered as outbreak periods, but only peak months of dengue every year.

Salim et.al. [9] evaluated machine learning models such as Decision Trees, Artificial Neural Networks, Support Vector Machines (SVM) and the Bayes Network for predicting dengue outbreaks in five districts with highest dengue incidence in Malaysia in 2013 to 2017. In their study, they used these machine learning models as classifiers to identify a value of a target variable which indicates whether or not there is a dengue outbreak during a week in a particular district. In particular, the dengue outbreak variable has a value of 1 if there is a dengue outbreak and 0 if there is no dengue outbreak during that week. Salim et.al. adopted the definition of the World Health Organization (WHO) for an outbreak. That is, an outbreak period is a period of time where the reported dengue cases is more that the sum of the moving average of three 4-week cases plus two standard deviations above the number of cases 4 weeks prior to the current period. Their results have shown that SVM performs best with the week feature as the best predictor variable.

A recent study by Chen et.al. performed outbreak detection in COVID occurrence through anomaly detection methods [4]. The outbreak detection problem was treated as a classification problem of manually labelled timestamps of COVID events as either anomalous or normal events. In their study, due to the lack of official reports on the specific dates for COVID outbreaks, anomalies were defined as sudden spikes in the of upward trends in the COVID positive cases. They have shown that anomaly detection models outperform baseline classification and clustering models in both precision and recall scores.

In another study by Jain et.al. [5], monthly dengue cases for each of the 50 districts of Thailand in 2008 to 2012 was used to forecast dengue cases using the GAM method and considered a threshold value to identify future outbreaks. They have shown that the model performs best when using lagged values of meteorological data, previous dengue cases and socioeconomic data as input to the GAM model.

In [6], outbreaks in dengue cases in provinces of Cambodia were identified based on detected abnormalities in the time series of dengue cases using the surveillance R-package and compared to the Bayesian approach. For the bayesian approach, they defined the upper limit of the cases count for a week to be the value that is greater that the 95th percentile distribution and the outbreak threshold for each province was based on the annual dengue peak in preceding years. In their detection approach, the seasonality in the epidemic was considered by considering only relevant data (i.e. data obtained in similar seasons). This approach however has a low robustness because it requires information on outbreak and non-outbreak periods.

3 Methodology

The machine learning classification models explored and tested in this study are the decision trees classifier, gradient boosting classifier, and the random forest classifier. The use of decision trees and gradient boosting classification models follows from the used method and the recommendation of Salim et.al to use

boosting algorithms [9]. Meanwhile, the random forest classifier is explored in this study as it has been shown to be the best machine learning model for predicting dengue incidence in Metro Manila, Philippines, in [3].

These three machine learning classification models will be tested for their performance in classifying outbreak and non-outbreak periods from 2015 to 2019 in the top 5 barangays with the highest cases yearly, namely Barangay Irisan, Barangay Camp 7, Barangay Bakakeng Central, Barangay Loakan Proper and Barangay Asin [1]. The best performing model will then be used to predict outbreak periods in these barangays in 2019 and 2020. The workflow for the clustering based approach and the classification approach for detecting possible dengue outbreaks is shown in Fig. 1.

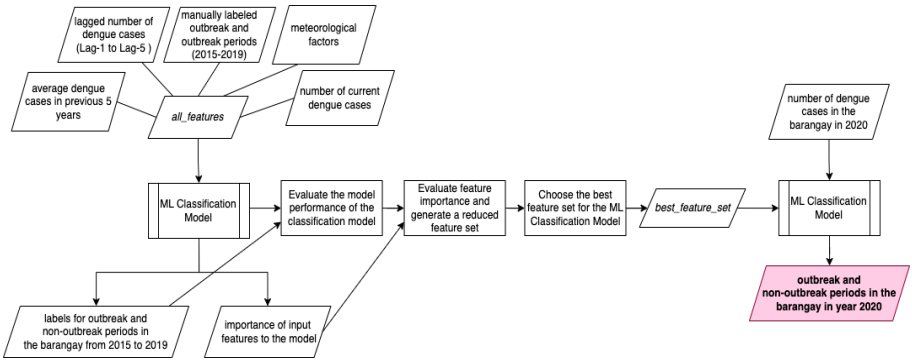


Fig. 1: Workflow for the detection of outbreak and non-outbreak periods of dengue in barangays using a machine learning classifier

3.1 Input Features to the Machine Learning Classifiers

The input features to the algorithm includes the average dengue cases in the previous 5 years, the lagged number of dengue cases, meteorological factors and the current number of dengue cases. These factors were chosen by combining features that have been shown to generate promising prediction models for incidence prediction model and outbreak prediction models in previous literature [9] [8] [5]. Particularly, the previous research in [9] and [5] have shown that including lagged number of dengue cases as input variable improves the model for prediction. Meteorological data such as temperature, rainfall and humidity have also been shown to be associated with dengue occurrence [8].

Putting all these features identified to be promising factors for outbreak prediction and dengue cases prediction in previous literature, the inputs to the ML classifiers for the outbreak detection are the average number of dengue cases (Ave), the lagged values of the dengue cases (Lag), the values describing the meteorological factors (MF) and the current number of dengue cases (C). Since

the behavior of dengue cases are also related to the factor of time, the month $m \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ in a year is also used as input to the model. In particular, one data input to the machine learning classification model is is a tuple x_t describing the following numerical values associated to a time period t:

$$x_t = (C, m, Ave, MF, Lag) \tag{1}$$

$$MF = (MaxT, MinT, MeanT, RH, WS, WD) \tag{2}$$

$$Lag = (Lag_1, Lag_2, Lag_3, Lag_4, Lag_5) \tag{3}$$

where each variable in the tuples MF and the Lag represent the following information at a particular time period t:

- RF is the amount of rainfall (in millimeters)
- $MaxT$ is the maximum temperature
- $MinT$ is the minimum temperature
- $MeanT$ is the average temperature
- RH is the relative humidity
- WS is the wind speed
- WD is the wind direction
- Lag_i is the number of dengue cases during the same time period i years before the current time period t

3.2 Training and Testing of the ML models

The ML classifiers will be trained on manually labeled data of outbreak and non-outbreak periods from 2015 to 2018, and tested on data from 2019 to 2020. In this study, the data are manually labeled as either an outbreak period or a non-outbreak period based on the average number of average dengue cases in the past five years prior to the current time period. In particular, a month of the year is labeled an outbreak month if the current number of dengue cases exceed the average of dengue cases recorded in the same month for the previous five years. Similarly, a week is labeled as an outbreak week if the number of cases during that week exceeds the average cases during the same week for the previous 5 years. Outbreak periods are labeled '`1`', while non-outbreak periods are labeled '`0`'. The machine learning classification models used are from the `sklearn` library of Python.

In training the ML models, a 5-cross fold validation technique is performed. In this technique, the ML classification algorithm is ran five times on the same set of input data and the training evaluation metric is determined by getting the average of the evaluation metrics obtained for each individual run of the classifier. The model with the best performance based on evaluation metrics is then used in classifying outbreak and non-outbreak periods in 2019 and 2020.

3.3 Evaluation metrics

The evaluation metrics used to evaluate the performance of the ML classification models are the prediction accuracy, precision, recall, F-score and Area Under the Curve (AUC) measures.

Prediction accuracy refers to the ratio of the number of correct predictions of the classification model out of the total number of predictions. It is given by the ratio of the sum of true positive (TP) and true negative (TN) predictions to the total number of predictions, that is:

$$PredictionAccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

On the other hand, precision measures the ratio of TP to the number of predicted positives, while recall is the ratio of TPs to the total number of actual positives, that is:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

where:

- TP (true positives) is the number of correctly classified outbreak periods
- TN (true negatives) is the number of correctly classified non-outbreak periods
- FP (false positives) is the number of outbreak periods that were classified as non-outbreak periods
- FN (false negatives) is the number of non-outbreak periods that were classified as outbreak periods

Finally, the F-score metric gives a combined measure of the precision and recall of the classification result and is given by:

$$F - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

A high recall value ensures that the classification model performs best in terms of predicting the actual outbreak periods. In this study, we choose the ML classification model with the highest prediction accuracy while prioritizing the model with a high recall and F-score.

3.4 Evaluation of feature importance

Similar to the study of Salim et.al. in [9], the importance of the input features to the model are evaluated in order explore which among the input features are most significant in detecting outbreak and non-outbreak periods. In this study, we use the `sci-kit learn` package in Python to generate plots to show the level

of importance of the feature inputs to the model. Features that are identified to be least important to the model is removed from the input feature set containing all features (*all_features*) to form the reduced feature set (*best_feature_set*). The reduced feature set is then used for predicting outbreak and non-outbreak periods in the year 2020.

4 Results and Discussion

Figure 2 plots the monthly average dengue cases in the past 5 years and the total dengue cases in 2015 to 2020 in the entire city. From this, a data input x_t is manually labeled as an outbreak month if the number of dengue cases at month t exceeds the average of cases in the previous 5 years, Ave. Out of the 72 months from 2015 to 2020, there are 53 non-outbreak periods and 19 outbreak periods in Baguio City. The 19 outbreak months occurred during in the following months: January, February, August to December of 2015, January to September of 2016, March 2018, and September to October of 2019.

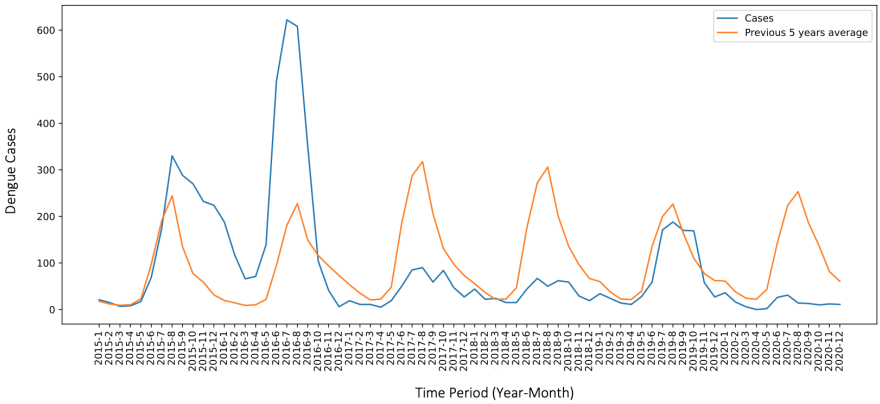


Fig. 2: Actual vs. average monthly dengue cases for the past five years from 2015 to 2020

Table 1 summarizes the model evaluation metrics of the machine learning classifiers used for predicting the outbreak months of dengue in Baguio City from 2015 to 2020. The precision results indicate that among the three ML models, the Random Forest, with a recall value of 1, was successful in identifying all actual outbreak months of dengue in the city from 2015 to 2020. The accuracy of the model at 90.90% and the precision of 66.66% indicates that out of all months that were classified as outbreak periods by the Random Forest, 66.66% were actual outbreak periods. Compared to the other two models, evaluation metrics

suggest that the Random Forest is the best model to be used for predicting outbreak months in Baguio City.

Table 1: Evaluation metrics of ML classifiers in identifying outbreak and non-outbreak months in Baguio City

Model	Prediction Accuracy	Precision	Recall	F-Score
Random Forest	90.90%	66.66%	1	0.8
Gradient Boosting	81.81%	0%	0	0
Decision Trees	72.72%	37.5%	0.75	0.5

For a more timely prediction of outbreaks, a weekly outbreak prediction was also done. Figure 3 plots the weekly average dengue cases in the past 5 years and the actual dengue cases in 2015 to 2020 for a total of 312 time periods. Similarly, an outbreak week is the week when the actual dengue cases exceed the average cases during the same week of the year in the previous 5 years. In this figure, there are 221 non-outbreak weeks and 91 outbreak weeks as tabulated in Table 2.

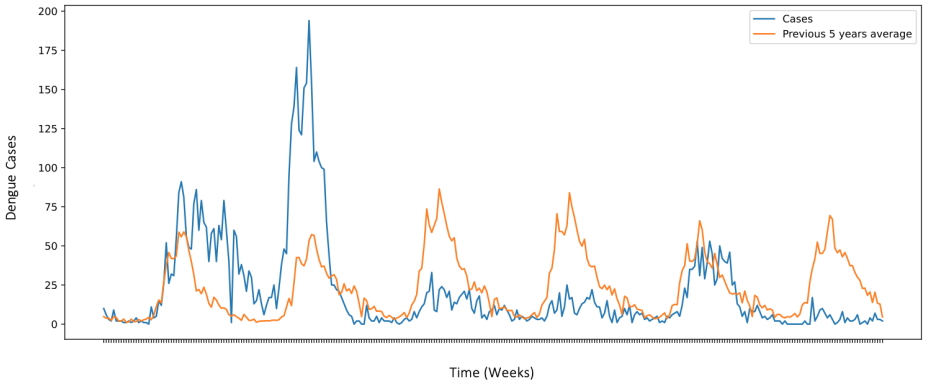


Fig. 3: Actual vs. average weekly dengue cases for the past five years from 2015 to 2020

Table 3 summarizes the model evaluation metrics of the machine learning classifiers used for predicting the outbreak weeks of dengue in Baguio City from 2015 to 2020.

Table 2: Outbreak weeks of dengue from 2015 to 2020 based on the previous 5 years average cases (Note: 1 year has 52 weeks from 0 to 51)

Year	Week Number
2015	Weeks 0,1, 4, 10, 12,13, 19,25,29,30-32 and 34-50
2016	Weeks 0-38 and 42
2017	Weeks 12 and 51
2018	Weeks 4,5,9,11 and 12
2019	Weeks 2,12,13,17,29,34-35,38-44 and 51

Table 3: Evaluation metrics of ML classifiers in identifying outbreak and non-outbreak weeks in Baguio City

Model	**Prediction Accuracy**	**Precision**	**Recall**	**F-Score**
Random Forest	81.81%	75%	0.50	0.60
Gradient Boosting	72.72%	0%	0	0
Decision Trees	86.36%	80%	0.667	0.72

The model evaluation metric of the ML classifiers suggests that a Decision Tree is the best model to be used for predicting outbreak weeks in Baguio City. The Decision Tree was able to identify 67% of the actual outbreak weeks, while 80% of the outbreak predictions of the Decision Tree are actual outbreak periods.

The results in Table 1 and Table 3 show that the Random Forest can accurately predict 90% of the expected outbreak months of dengue in the city. This result shows that the Random Forest may also be used as a classification model for outbreak detection in addition to being the best machine learning model for predicting dengue incidence as shown by a study by Carvajal et.al. [3]. On the other hand, the Decision Tree classifier is most accurate in predicting outbreak weeks in the city at 86.36% prediction accuracy.

Based on the average of cases in the past 5 years, the number of dengue outbreak months from 2015 to 2018 in each of the barangay is 27, 26, 29, 27 and 26, respectively. Table 4 summarizes the model statistics of training the random forest model for each barangay. All models obtained the same precision and recall statistics but the Random Forest model obtained the highest accuracy in predicting monthly outbreaks in all 5 barangays.

The random forest classifier is then used to predict outbreak months of dengue from 2019 to 2020 in each barangay. The results of the final prediction of outbreak periods were obtained using a bagging technique of the prediction results of 5 runs of the random forest model. Prediction are shown in Figure 5 and Figure 6 wherein the predicted dengue outbreak months are indicated by the green diamonds.

Table 4: Training evaluation metrics for a random forest classifier in identifying outbreak and non-outbreak months in barangays with highest dengue cases yearly

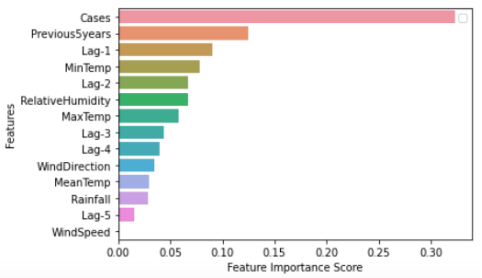
Barangay	Accuracy	AUC	Precision	Recall	F-Score
Irisan	100%	0.8545	80%	70.91%	0.7412
Camp 7	99.58%	0.8789	93.33%	77.39%	0.8219
Bakakeng Central	100%	0.8977	92.5%	84.34%	0.8766
Loakan Proper	100%	0.8286	80%	65.71%	0.6889
Asin	100%	0.8	60%	60%	0.6

Generally, the random forest classifier was successful in detecting most of the expected outbreak months in most of the barangays. Majority of the expected outbreaks during the peak months in 2019 was detected by the classifier, except for the two months outbreak in Barangay Asin. Moreover, there are also predicted outbreak months that are not actual outbreak months. Nonetheless, although this would entail a false alarm to an outbreak, it is more favored than predicting that an outbreak month is not an outbreak month. Table 5 summarizes the model statistics of the random forest in predicting outbreak months in 2019 to 2020 for 5 barangays.

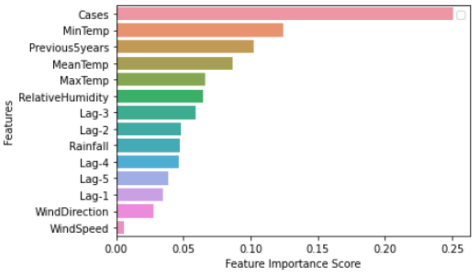
Table 5: Testing evaluation metrics for a random forest classifier in identifying outbreak and non-outbreak months in barangays with highest dengue cases yearly

Barangay	Accuracy	AUC	Precision	Recall	F-Score
Irisan	78.33%	0.6758	30.54%	70.91%	0.7412
Camp 7	90.83%	0.7143	60%	46.67%	0.5143
Bakakeng Central	82.5%	0.8048	47.5%	73.33%	0.5591
Loakan Proper	71.67%	0.6278	47.5%	40%	0.3714
Asin	79.17%	0.6911	53.56%	42.22%	0.4450

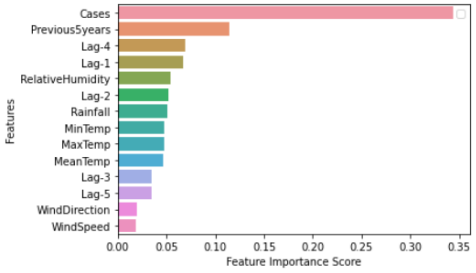
Following this result, we attempt to improve the model prediction by testing the model's performance when only the most important input features are used as input to the model. Figure 4 shows the feature importance of the input features used in the random forest classifier. It shows that the feature importance of the input features to the random forest classifier vary for each barangay. Nonetheless, the average of cases in the previous years, Ave and temperature are seen to have high importance in all barangays. The relative humidity RH also shows high importance in most of the barangays. This result is consistent with the finding



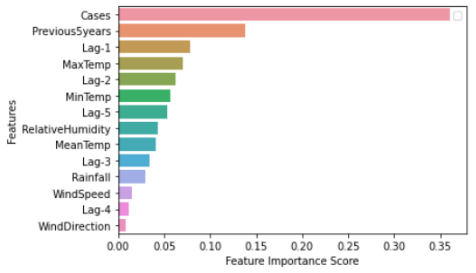
(a) Barangay Irisan



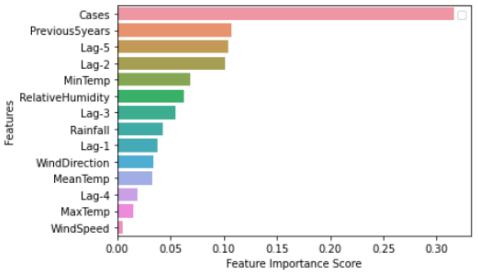
(b) Barangay Camp 7



(c) Barangay Bakakeng Central



(d) Barangay Loakan Proper



(e) Barangay Asin

Fig. 4: Importance of input features for predicting outbreaks using the random forest classifier

of Marigmen et. al. in [8] wherein humidity and precipitation was found to be highly associated to the dengue cases in Baguio City. This then suggests that humidity may also be considered as a factor for dengue outbreak prediction in the barangays of Baguio City. Moreover, the windspeed, WS, is the least important input feature to the ML classifier, thus indicating that this feature may be omitted in designing a model for predicting dengue outbreak months in the barangays.

The input features to the random forest classifier for all barangays is reduced to the following set of features with the month, m included:

$$Feat = [Ave, RH, MeanT, MaxT, MinT, RH, Lag1, Lag2, Lag3, m] \quad (8)$$

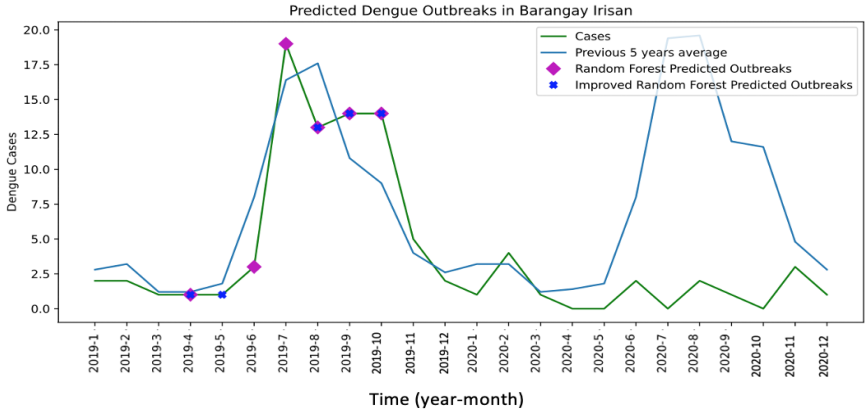
Table 6: Testing results of the random forest classifier using a reduced feature set

Barangay	**Accuracy**	**AUC**	**Precision**	**Recall**	**F-Score**
Irisan	75.00%	0.621	40%	40%	0.4
Camp 7	79.20%	0.738	33.33%	66.70%	0.444
Bakakeng Central	87.50%	0.929	50%	100%	0.667
Loakan Proper	79.20%	0.694	60.00 %	40%	0.545
Asin	83.30%	0.778	100%	55.60%	0.714

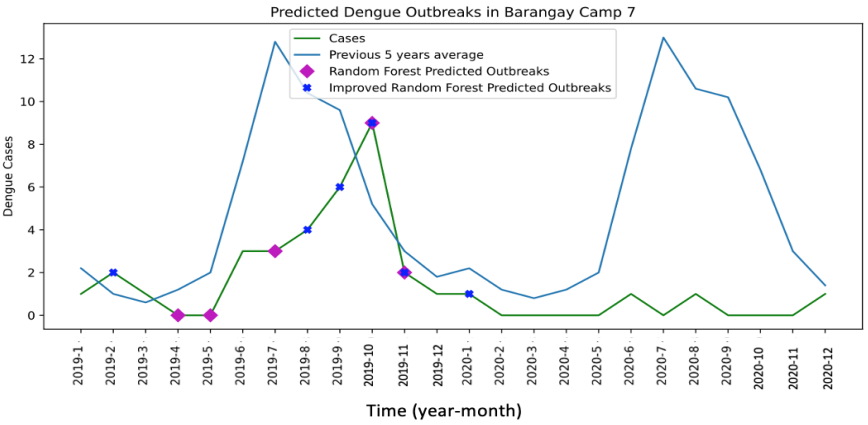
The green diamonds in Figure 5 and Figure 6 indicate the shows the predicted dengue outbreak months by the Random Forest classifier and Table 6 presents the model prediction statistics.

5 Conclusion

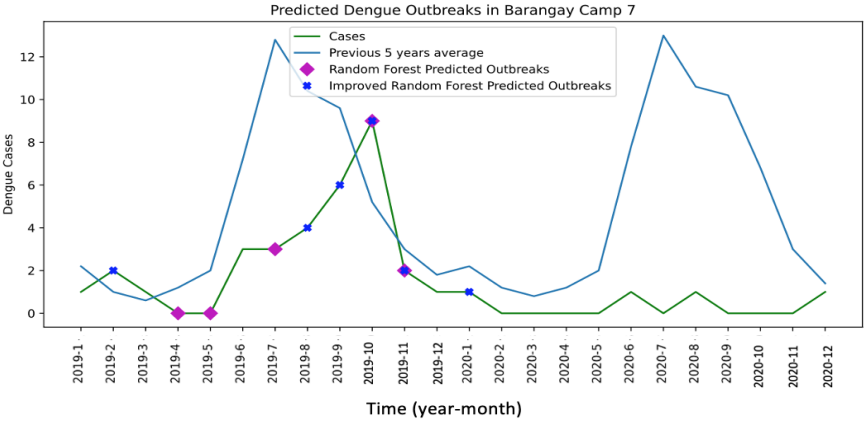
The detection of dengue outbreaks in local communities in Baguio City, Philippines, can be thought of as a classification problem using machine learning classification models such as the Random Forest classifier. Similar to previous studies showing the correlation of the rise of dengue cases to the humidity and temperature, these factors, including the lagged dengue cases has shown promising classification results of outbreak and non-outbreak periods of dengue. It is highly recommended that the Random Forest classifier be tested in predicting outbreak periods in the next years subject to the availability of dengue data and meteorological data. Employing forecasting methods to forecast the number of dengue cases before performing the outbreak prediction is also recommended.



(a) Barangay Irisan

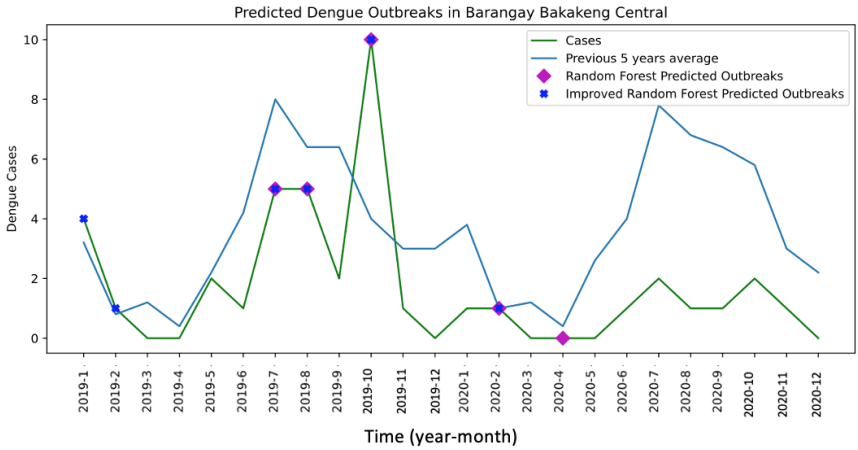


(b) Barangay Camp 7

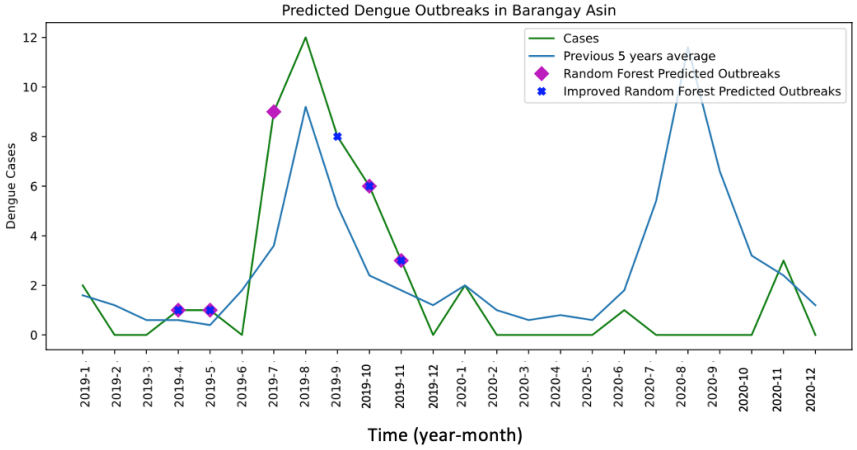


(c) Barangay Loakan Proper

Fig. 5: Predicted outbreak months in the top barangays in Baguio City from 2019 to 2020



(a) Barangay Bakakeng Central



(b) Barangay Asin

Fig. 6: Predicted outbreak months in the top barangays in Baguio City from 2019 to 2020 (continued)

References

1. Addawe, J.C., Caro, J.D.L., Juayong, R.A.B. (2023). Machine Learning Methods for Modeling Dengue Incidence in Local Communities. In: Krouska, A., Troussas, C., Caro, J. (eds) Novel & Intelligent Digital Systems: Proceedings of the 2nd International Conference (NiDS 2022). NiDS 2022. Lecture Notes in Networks and Systems, vol 556. Springer, Cham. https://doi.org/10.1007/978-3-031-17601-2_38
2. Baharom, M., Ahmad, N., Hod, R., and Manaf, M. Dengue early warn- ing system as outbreak prediction tool: A systematic review. Risk Management and Healthcare Policy Volume 15 (05 2022), 871–886.
3. Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D., and Watanabe, K. Machine learning methods reveal the tem- poral pattern of dengue incidence using meteorological factors in metropolitan manila, philippines. BMC Infectious Diseases 18 (2018).
4. Chen, Y. Detecting covid-19 outbreak with anomalous term frequency.
5. Jain, R., Sontisirikit, S., Iamsirithaworn, S., and Prendinger, H. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC Infectious Diseases 19 (03 2019).
6. Ledien, J., Souv, K., Leang, R., Huy, R., Cousien, A., Peas, M., Froehlich, Y., Duboz, R., Ong, S., Duong, V., Buchy, P., Dussart, P., and Tarantola, A. An algorithm applied to national surveillance data for the early detection of major dengue outbreaks in cambodia. PLOS ONE 14, 2 (02 2019), 1–11.
7. Marigmen, J. L. D. C., and Addawe, R. C. Analysis on the onset of dengue outbreaks in Baguio city. AIP Conference Proceedings 2423, 1 (11 2021). 070012.
8. Marigmen, J. L. D. C., and Addawe, R. C. Forecasting and on the influence of climatic factors on rising dengue incidence in baguio city, philippines. Journal of Computational Innovation and Analytics (JCIA) 1, 1 (2022), 43–68.
9. Azam, N., Salim, M., Yap, B., Reeves, C., Smith, M., Fairos, W., Wan Yaacob, W. F., Mudin, R. M., Dapari, R., Fatin, N., Fatihah, F., and Haque, U. Prediction of dengue outbreak in selangor malaysia using machine learning techniques. Scientific Reports 11 (01 2021).