Dengue Risk Mapping from Geospatial Data Using GIS and Data Mining Techniques

https://doi.org/10.3991/ijoe.v16i11.16455

Benjawan Hnusuwan, Siriwan Kajornkasirat, Supattra Puttinaovarat (✉)
Prince of Songkla University, Surat Thani, Thailand
supattra.p@psu.ac.th

Abstract—Dengue fever is a major public health problem and has been an epidemic in Thailand for a long time. Therefore, there is a need to find a way to prevent the disease. This research aimed to explore the important factors of dengue fever, to study the factors affecting dengue hemorrhagic fever in Surat Thani Province, and to map the potential outbreak of dengue fever. Collecting patient information was done including, Rainfall, Digital Elevation Model (DEM), Land Use and Land Cover (LULC), Population Density, and Patients in Surat Thani Province, which was analyzed using data mining techniques involving analysis using 3 algorithms comprising Random Forest, J48, and Random Tree. The correct result is Random Forest since the accuracy of the data is 96.7 percent followed by J48 with accuracy of 95.9 percent. The final sequence is Random Tree with accuracy of 93.5 percent. Then, using the information can be displayed through ArcGIS program to see the risk points that are compared to the risk areas that have been previously done. The results can be very risky in Mueang District, Kanchanadit District, and Don Sak District, corresponding to the information obtained from the Public Health Office and the risk map created from the patient information.

Keywords—Dengue risk mapping, Geospatial data, Data mining, GIS.

1 Introduction

Dengue fever is identified as a disease caused by the Dengue virus. There are around 2.5 billion people worldwide who have been infected with dengue fever [1-4]. Outbreak can occur in the rainy season from Lai mosquitoes, which like to be active in the daytime in homes and schools that have dense populations. Areas with water sources have Lai mosquitoes because they have water that is watery and clear. After the rain, because the temperature and humidity are suitable for breeding in other seasons, the prevalence of striped mosquitoes decreases slightly [5]. Most will occur in tropical countries [6-7]. Dengue infection can result in very mild illness or death. The symptoms often begin with headaches, muscle pain, and bone pain, which can be divided according to the 3 phases of the illness, namely fever phase, shock phase and recovery period [8-9].

Since 1953-1964, dengue fever has spread in many countries in Southeast Asia and Pacific Asia, namely the Philippines, Thailand, Vietnam, Singapore, and Kolkata in

India [10]. Thailand has reported dengue outbreaks for more than 60 years. At present, moisture affects the epidemic of dengue fever, which starts from June to August. There will be more severe results when the temperature exceeds 24 ° C to 30 ° C [11], but it cannot spread if the temperature is below 16 ° C [5]. Dengue fever can cause an outbreak that can spread throughout the country, into every province and district. The distribution of the disease has changed over time. According to the report of the dengue hemorrhagic fever situation in 2019, there have been 715 cases of dengue fever (DHF, Dengue shock syndrome: DSS), increasing from 621 last weeks. Sick 1.08 per hundred thousand people there have been more reports of dengue fever than in 2017 at the same time of 2.6 times the spread of dengue fever. It was found that the southern region had the highest rate of illness at 3.14 per hundred thousand people. The number of 294 patients, followed by the central region, 1.04 patients per hundred thousand people, 99 cases and the Northeast region, with a rate of 0.41 per 100,000 Population Density, 89 patients, respectively [12].

Methods found from the literature review include using experts to give weight and score values (Weighted multi-criteria: WMC), which is an easy way to determine which data layer or factor is less important. By assigning numbers such as 10 (most important) 9 8 7 ... sorting less important, which in the data layer of each factor the same. Using experts, each branch will be an assistant in scoring, after which the values obtained are multiplied and added together according to general suitability equations [5] [13-14]. The method described above is delayed because it requires contact with an expert, Forms for weighting and rating values are created. When changing the data in the analysis, the above steps are repeated, meaning it is not easy to work. However, in the case that we use data mining techniques when changing data or analysis factors. The results can be analyzed through the model and confidence obtained. This technique is fast and convenient to be able to change any factors. To suit the research immediately, there are many statistical methods used such as Multi-criteria decision analysis [13], Ordinary least squares (OLS), Generalized linear mixed model (GLMM), Naïve Bayes [15] Multi regression, k- means Clustering [16] etc. Also, there are different levels of risk including very risky, risky, moderate risk, low risk and not risky [17] 4 levels are very risky, risky, moderate risk and not risky [5] or 3 levels is very risky, risky and not risky.

From the research study, it was found that land-use factors, seasons, rainfall, temperature, humidity [6][15][18], height, river, Population Density [13][17] and applying Geographic Information System (GIS) tools to predict or define risk areas by linking, analyzing, processing and displaying the relationship of data [19-21]. From the data, the introductory researchers are interested in studying and processing the data. By bringing annual data about dengue patients in Surat Thani Province Between January 2018 and September 2018, the research has 3 main objectives: 1) to explore the important factors of dengue fever 2) to study the factors affecting dengue hemorrhagic fever in Surat Thani Province, and 3) to map the outbreak of dengue fever, which will be applied to plan and prevent the spread of dengue fever.

2 Materials and Methods

2.1 Study area

Surat Thani is a province in the upper southern region. It has the largest area in the south and the 6th largest in Thailand. It has the 59th highest density population in the country. Surat Thani Province is located on the eastern side of the southern region. There is a variety of terrain, including flat terrain, coastline, plateau, as well as mountainous terrain. Thailand has reported dengue fever for over 60 years. Currently, dengue fever is spread throughout the country in every province and district. The spread of the disease has changed over time. Surat Thani is a province that ranks first. That is found in the most dengue patients from 1 January to 29 September 2018, a total of 939 patients were added from the previous year and are likely to increase in 2019. From the forecasting report Dengue fever 2019 by the Department of Disease Control, including Bureau of Communicable Diseases by Insects, Bureau of Epidemiology, Office of Disease Prevention and Control 1-12 and, Urban Disease Prevention and Control Institutions stated that the southern region is a risk area for dengue fever epidemic. The provinces that are expected to have the most severe outbreak include Nakhon Si Thammarat, Krabi and Surat Thani. In Surat Thani, the number of patients is expected to reach 1,140 [12]. Figure 1 shows the boundaries of the study area, which is every district in Surat Thani Province, consisting of 131 districts together.

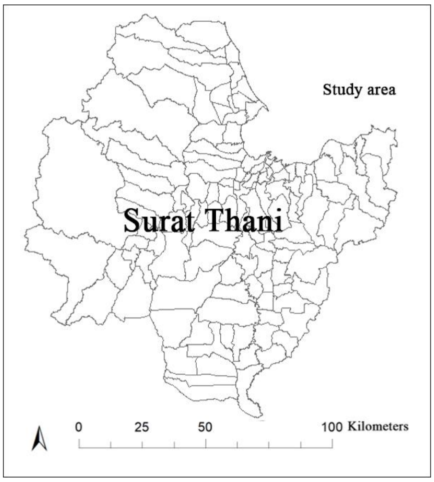

Fig. 1. Study area within Surat Thani province, Thailand

2.2 Abbreviation

Table 1 reveals the abbreviation that we need to use along the research paper. Though, these abbreviations will be defined in the text.

Table 1. Abbreviations used in the research paper and its meanings

Abbreviation	Description
°C	Degree Celsius
DEM	Digital Elevation Model
DHF	Dengue Hemorrhagic Fever
GDP	Gross Domestic Product
GIS	Geographic Information System
GLMM	Generalized Linear Mixed Model
IDW	Inverse Distance Weight
LULC	Land Use and Land Cover
m	Meter
mm	Millimeter
OLS	Ordinary least squares
RMSE	Root Means Square Error
WHO	World Health Organization
WLC	Weighted Linear Combination
WMC	Weighted multi-criteria

2.3 Data collection

Study relevant factors from various researches both in Thailand and abroad then bring to compare to see what factors are appropriate and popular in the analysis for the mapping of dengue fever. Table 2 shows a comparison of factors such as season, Rainfall, Temperature, Humidity, DEM, Slope, LULC, River, Water Source, Gross Domestic Product (GDP), Economy, and Population Density. The factors are chosen from the highest total score of three, i.e. 6, 5 and 4, which has all 6 factors from 12 factors including Rainfall, Temperature, Humidity, DEM, LULC, and Population Density.

When all the factors data is collected by requesting information from various agencies, the data that we use is data in 2018 for Surat Thani Province. Both spatial data and quantitative information exist as follows. Rainfall data, Temperature data, Humidity data is the data obtained from the Meteorological Department [22] DEM data, LULC data are the information obtained from Land Development Department [22]. Population Density data is information obtained from the National Statistical Office [23], while Patient Data is information obtained from the Surat Thani Provincial Health Office [24].

Table 2. Literature review of DHF factor comparing

Order	**Research**	**Factors Data**											
		Season	***Rainfall***	***Temperature***	***Humidity***	***DEM***	***Slope***	***LULC***	***River***	***Water source***	***GDP***	***Economy***	***Population Density***
1	Zambrano et al., 2017 [6]		✓	✓	✓	✓							
2	Agarwal et al., 2018 [16]		✓	✓	✓	✓							✓
3	Yue et al., 2018 [25]			✓	✓			✓			✓		✓
4	Liu et al., 2018 [26]		✓		✓					✓		✓	
5	Minale and Alemu, 2018 [18]		✓	✓		✓	✓	✓					
6	Fuller et al., 2014 [13]					✓			✓				✓
7	Sammatat et al., 2013 [15]	✓	✓	✓									✓
8	Somard et al., 2015 [19]							✓					✓
9	Sermkarndee et al., 2016[5]		✓	✓	✓		✓		✓				

2.4 Data preprocessing

The factors data were derived from various sources and different in units. To ensure homogeneity, each factor was prepared by first rate its raw value into discrete level as in Table 3. The sampling was made uniformly 1000 instances (points). In the raining and predicting in this research, the K-fold Cross Validation was employed as a metric to evaluate the merit of each configuration. The data ware divided equally in to K groups. For each round, one group was selected for training the model, while the remaining K-1 groups were used testing, from which the results were compared with the actual result, and therefore the predicting error was calculated. In order to minimize data dependent biases, the process was repeated K times, each with training data iterating through a different group of datasets. In this research, the value K was set to 10. To address over-fitting issues 10-fold cross validation was used to assess all abovementioned data mining algorithms, in the experiments.

2.5 Spatial analysis

ArcGIS is used for data pre-processing, spatial analysis and visualization [27]. Starting, we will take the information on the factors that we have chosen, including Rainfall, Temperature, Humidity, DEM, LULC. Population Density and the number of Patients. Come to Interval - Ratio Level divide the data group or separate the categories to indicate the level of difference between groups or categories. Use the comparison of characteristics in each factor. Once the information of all factors has been obtained, this experiment, use 1000 points to be able as a representative for each area to create a map for dengue fever monitoring. However, more than 1000 points have been used in testing. Which in analyzing the data for use in real situations, the size of datasets can be increased without decreasing the accuracy. The class which will be used to see the correctness of the model is patient data. Used to Interval - Ratio Level data to see which areas are at risk of outbreaks of disease-causing patients to occur, which will be divided into 4 classes include, 4 layers, shown on the map to see which areas are at risk of the dengue outbreak include, Level 4 high risk is red area, Level 3 moderate risk is orange area, Level 2 mild risk is light green and Level 1 low risk is dark green.

Rainfall for the amount of rainfall data used in this research, the monthly rainfall is used May to September of 2018. Surat Thani has set rain all year round, causing a puddle of floods. Rainfall is, therefore, a very important factor used in the analysis, which has a total of 9 stations including Chaiya, Bannasan, Kanchanadit, Phunphin, Tha-Chana, Phanom, Khun-Thale, Surat Thani Rubber Research Center Station, and Surat Thani City Station. Therefore, bringing all the data to estimate the value during (Interpolate) by IDW (Inverse Distance Weight) technique is estimation by randomly sampling each sample point from a position that can affect the cells to be estimated. This will have less impact according to the distance [28-29].

Figure 2 by the red area is the area with the highest amount of rainfall. This research selected using IDW precipitation estimation because this method provides the least average error, especially the monthly and annual rainfall estimation. In addition, the estimation of water values in this research found that the IDW method of precipitation is able to distinguish the rainfall data for each area the best.

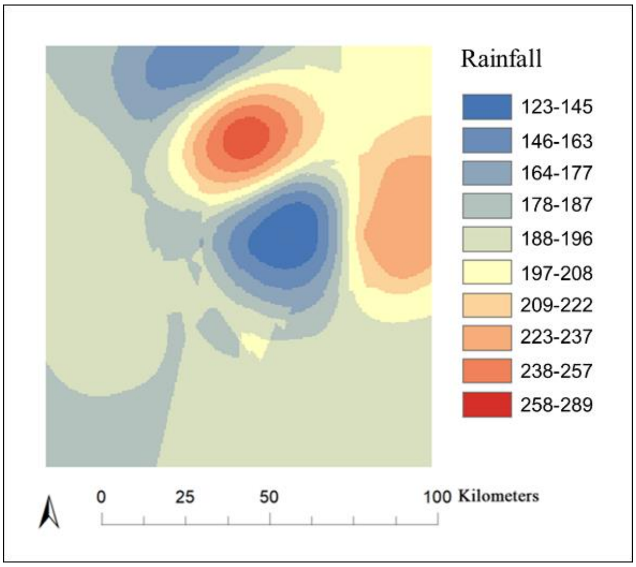

Fig. 2. Rainfall-Level Map in Surat Thani Province

DEM identify the source of mosquito problem (i.e. vector breeding areas). It was found that more than ninety percent of the case samples were in the 'High' and 'Very High' categories [30]. This research is divided into 10 levels of altitude, with the height of the area starting at 200 meters from the mean sea level. The west side of the area begins to rise gradually until the east side of the area looks like a ridge.

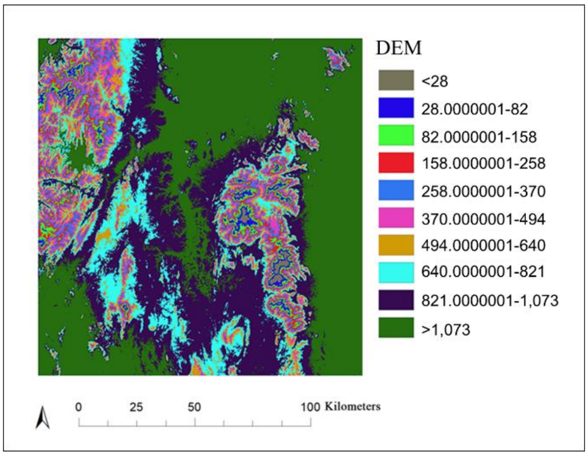

Fig. 3. DEM-Level Map in Surat Thani Province

Population Density is a measure of the Population Density in a given area, depending on the sample chosen to be surveyed. We choose the Population Density in the residential area, which the data used from the National Statistics Department, as shown in Figure 4.

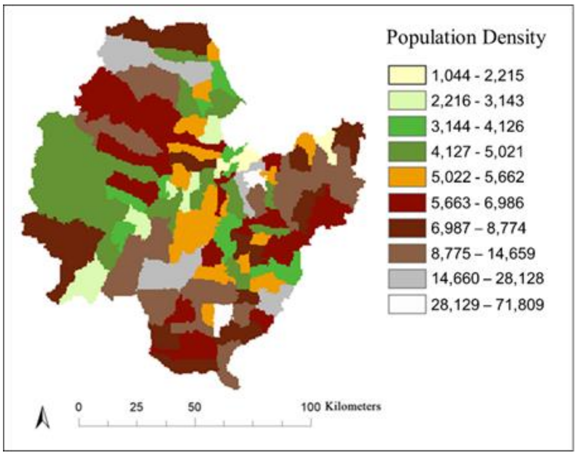

Fig. 4. Population Density-Level Map in Surat Thani Province

Land use and land covering data in Surat Thani Province can be classified into 6 groups, as in Figure 5, as follows: Green area is the Other utility space (O), Light blue area is the Water source area (W), Yellow area is the Agriculture (A), Pink area is the Mix can't distinguish what type of area (M), Red area is the Urban community area (U), and Blue is the Forest area.

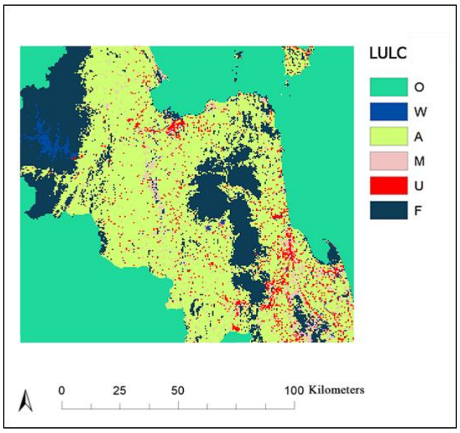

Fig. 5. LULC Map in Surat Thani Province

Patient Data used in this research comprised district-level patient data in 2018, with a total of 939 patients in that year. It was found that the top 5 patients with dengue fever were Ma Kham Tia 96 cases, Bang Kung 32 cases, Talad 29 cases, Tha Thong Mai 27 cases, and Ban Na San 22 cases. The study will divide the data into 4 groups as in Figure 6: dark green is a group of 0-5 patients, Light green is a group of 6-16 patients, Orange is a group of 17-37 patients, and red is Group of 38-102 patients.

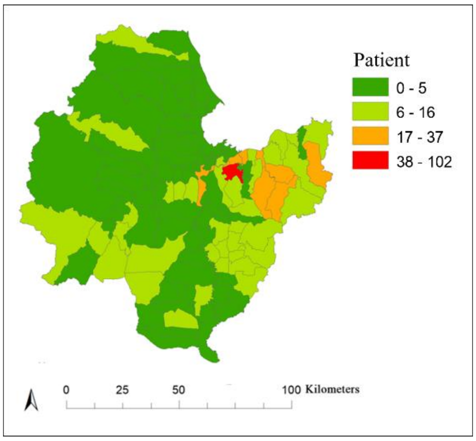

Fig. 6. Patient-Level Map in Surat Thani Province

This research uses data mining techniques to analyze the accuracy of the data because it is a fast and searchable method. Relationships that are hidden in that data set easily [31-32]. It does not require the WLC method to analyze the suitability of the factor because it is a delayed and multi-step because of the need to apply the scoring criteria obtained from the classification of each factor (Rating) in the analysis together with the weight factor obtained from experts. That is to say, the WLC method must evaluate the value multiplied by the weighting of each factor and find the sum. After that, the method of dividing into a class of security is used in analyzing the level of suitability. In the case of this method, if there is a change or addition of other factors incoming, they must be considered by all new experts. In this research, we propose a method for analysis using data mining techniques with data obtained from Interval - Ratio Level to lead the analysis shown in Table 3: Rainfall, DEM, LULC, Population Density, and Patients factor.

Table 3. Factors affecting the occurrence of dengue fever by using data mining techniques

Factor	Data Value		Rating
Rainfall	258-289	mm	10
	238-257	mm	9
	223-237	mm	8
	209-222	mm	7
	197-208	mm	6
	188-196	mm	5
	178-187	mm	4
	164-177	mm	3
	146-163	mm	2
	123-145	mm	1
DEM	>1,073	m	10
	1,073-821.0000001	m	9
	821- 640.0000001	m	8
	640-494.0000001	m	7
	494-370.0000001	m	6
	370-258.0000001	m	5
	258-158.0000001	m	4
	158-82.0000001	m	3
	82-28.0000001	m	2
	<28	m	1
LULC	F (Forest)		6
	U (Urban)		5
	M (Miscellaneous)		4
	A (agriculture)		3
	W (Water)		2
	O (Others)		1
Population Density	71,809 - 28,129		10
	28,128 - 14,660		9
	14,659 - 8,775		8
	8,774 - 6,987		7
	6,986 - 5,663		6

Factor	Data Value		Rating
	5,662 - 5,022		5
	5,021- 4,127		4
	3,144 - 4,126		3
	3,143 - 2,216		2
	2,215 - 1,044		1
Patients	38-102		4
	17-37		3
	6-16		2
	0-5		1

When dividing all Interval - Ratio Level, we will export the data in the form of .pdf files leading to the process of data analysis to see the risk areas and random points that we have randomly matched. We will verify the accuracy of data analysis using data mining techniques.

2.6 Data mining

The learning and testing process were performed using WEKA. However, other tools such as Python, Scikit-Learn, R Studio and RapidMiner can be used to analyze. The data mining process consists of sub-work flows that turn raw data into knowledge. The steps are shown in Figure 7: Data Cleaning is a procedure for eliminating unrelated data, while Data Integration is the process of combining data with multiple sources into one set of data. Data Transformation is a data conversion procedure that is suitable for use, while Data Reduction is the process to reduce the complexity of data [31-32]. We choose to use the Decision Tree Algorithm as it is the algorithm that is widely known and suitable for solving complex problems. It is a model that is easy to understand, consisting of Random Forest, J48 and Random Tree. The reason for choosing the 3 algorithms is Random Forest is popular, has excellent performance and is accurate in classification tasks. It even outperforms its counterparts such as discriminant analysis, neural networks and support vector machines [33]. For J48, the study found that it is applied to other fields. It is not related to health, but is used in the analysis of credit, in which the research is compared with other algorithms. The results obtained from the analysis are better than Naïve Bayes and PART [34], so it is chosen to apply for health research. Random Tree is an algorithm that is rarely seen in analysis. It is interesting that the results will be different from the algorithm that is popularly used and will give good results in health research.

All 3 algorithms are used in instructional learning. It is a learning method that is not very complicated builds classification or regression models are in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets, while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor, called a root node. Decision trees can handle both categorical and numerical data [35-37]. Its work has to be divided into two parts: training data and testing data. Training data will be a guide for teaching the machine to learn the data that is analyzed first. Then the information that we want is added to put in to see how the data is correct and how reliable the models are in our analysis to predict which areas are at risk.

We use patient data for consideration together with the other 4 factors, which are Rainfall, DEM, LULC, and Population Density. When the results of the analysis are required, Pattern Evaluation is the process of evaluating patterns obtained from data mining. Knowledge Representation is the process of presenting knowledge that has been discovered [31-32].

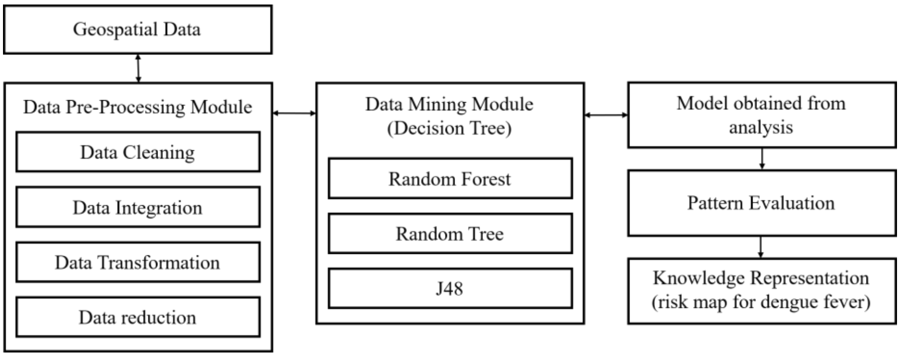

Fig. 7. Workflow for Data Mining Technique

Random forest: Random forest is a popular machine learning procedure which can be used to develop prediction models [38]. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set [39]. The input of each tree is sampled data from the original dataset. In addition, a subset of features is randomly selected from the optional features to grow the tree at each node. Each tree is grown without pruning. Essentially, a random forest enables a large number of weak or weakly-correlated classifiers to form a strong classifier [40-41].

$$Err(\varphi_{\mathcal{L}}) = E_{X,Y}\{L(Y, \varphi_{\mathcal{L}}(X))\} \tag{1}$$

Similarly, the expected prediction error of $\boldsymbol{\varphi_{\mathcal{L}}}$ at $\boldsymbol{X} = \boldsymbol{x}$ can be expressed as

$$Err(\varphi_{\mathcal{L}}(X)) = E_{X|Y=x}\{L(Y, \varphi_{\mathcal{L}}(X))\} \tag{2}$$

In regression, for the squared error loss, this latter form of the expected prediction error additively decomposes into bias and variance terms, which together constitutes a very useful framework for diagnosing the prediction error of a model. In classification, a similar decomposition is more difficult to obtain for the zero-one loss. Yet, the concepts of bias and variance can be transposed in several ways to classification, thereby providing comparable frameworks for studying the prediction errors of classifiers [42].

In the proposed model, Random Forest consisted of four parameters, i.e., numClasses, maxDepth, numFeatures and Iteration, as shown in Table 4.

Table 4. Random Forest employed in this study and its parameter settings

Parameter	Value
numClasses	4
maxDepth	0 (Unlimited)
numFeatures	0
Iteration	500

Random tree: Random Trees is a supervised Classifier; it is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree. In a standard tree, every node is split using the best split among all variables and outputs the class label that received the majority of "votes" [43]. This method is called Random Trees because you are actually classifying the dataset a number of times based on a random sub-selection of training pixels, thus resulting in many decision trees. To make a final decision, each tree has a vote. This process works to mitigate over-fitting. Random Trees is a supervised machine learning classifier based on constructing a multitude of decision trees, then choosing random subsets of variables for each tree [44].

Decision tree (J48): Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm finds out the way the attributes-vector behaves for a number of instances. Also, the classes for newly-generated instances are found on the basis of the training instances. This algorithm generates the rules for the prediction of the target variable. With the help of the tree classification algorithm, the critical distribution of the data is easily understandable [45]. This process uses "Entropy", which is a measure of the data disorder. Entropy is calculated by:

$$\boldsymbol{Entropy(\vec{y}) = -\sum\nolimits_{j=1} \frac{|y_i|}{|\vec{y}|} log\left(\frac{|y_i|}{|\vec{y}|}\right)} \tag{3}$$

$$\boldsymbol{Entropy(j|\vec{y}) = \frac{|y_i|}{|\vec{y}|} log\left(\frac{|y_i|}{|\vec{y}|}\right)}$$

$$\boldsymbol{Gain(\vec{y}, j) = Entropy\left(\vec{y} - 45 -=\right.} Entropy\left(\dot{\boldsymbol{j}}|\vec{\boldsymbol{y}}\right)\Big) \,|$$

The objective is to maximize the Gain, dividing by overall entropy due to split argument $\vec{\boldsymbol{y}}$ by value j.

2.7 Accuracy assessment

This paper employed standard accuracy metrics which were Accuracy, Kappa, Root Mean Square Error (RMSE). Particularly, RMSE were defined as follow:

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(x_i - x_i^{'})^2} \tag{4}$$

where n was the number of data samples, x_i and x_i^' were the actual and predicted values, respectively.

3 Result

The application of various techniques to study the factors affecting the outbreak of dengue fever in Surat Thani Province can be divided into 2 main parts: Spatial Analysis and Data Mining Analysis. By compiling a variety of data such as Rainfall, DEM, LULC, Population Density, and Patients, which is the data in the year 2018, all the data is taken through the Interval - Ratio Level to determine the relationship between each factor. The data obtained from the analysis using data mining techniques such as Random Forest, Random Tree, and J48 were utilized to find the most suitable model for creating dengue risk maps.

Figures 8, 9, and 10 show the risk points on the Surat Thani map, which divides the risk areas for dengue fever into 4 levels including no risk (green), low risk (yellow), moderate risk (orange) and high risk (red). Due to the model obtained not being very different, the risk point is similar. This can be explained as follows: Very risk areas have a lot of distribution in Surat Thani, which includes Makham Tia Sub-district, which is an area that is high from sea level with a large urban area and many people. For moderate risk areas, there is a lot of distribution in Kanchanadit District, including Pa Ron, Chang Sai, Krut Sub district and also found distribution in Phunphin district including, Tha Kham district also. These two areas are elevated high away from sea level. Most of the usable areas are urban community, agricultural, and forest areas, consistent with the data we have studied because more space is used for living and agriculture. This part will stimulate the amount of waste and additional equipment for daily use. The usual breeding places are the roof gutters, flower pots, flower pot plates, and roadside drains. The breeding places may also be in unexpected places such as plant axils, tree holes, air-conditioners, canvas sheets, and discarded receptacles in the area. There is a risk of outbreak for dengue fever [46]. Low risk and no risk areas are not too far from sea level. Most of these areas are agricultural areas, forests, and water resources, meaning there is not high density.

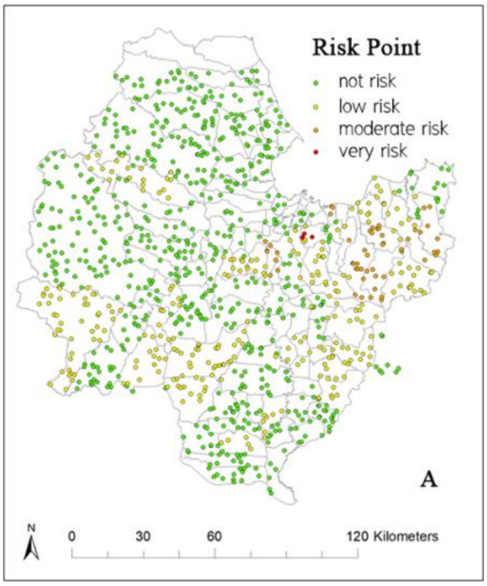

Fig. 8. Risk analysis result from the Random forest model.

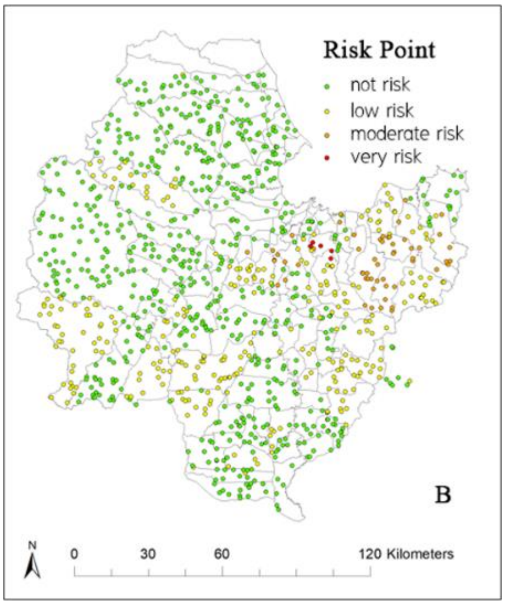

Fig. 9. Risk analysis result from the Random Tree model.

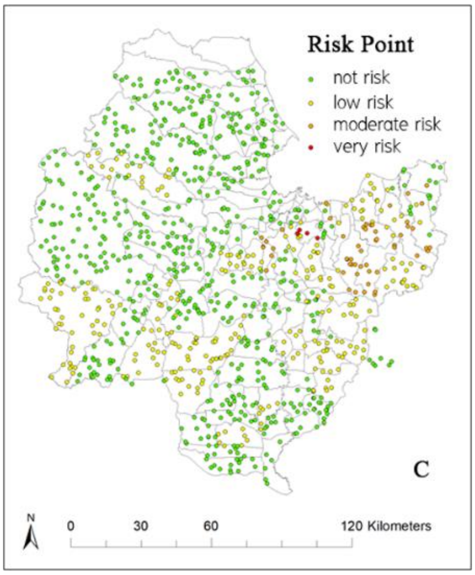

Fig. 10. Risk analysis result from the Decision Tree (J48) model.

Figures 11, 12, 13 show the validation of the model from Random Forest, Random Tree, and J48, which is divided into 2 colors including the True point (green) and the False point (red). Figure 11 shows model checking with data to there are any points on the map that the model has examined correctly or which ones are wrong in Random Forest from data 1000 points, correct 967 points and error 33 points. The point where the model analyzes has the most data errors. The area has agriculture, urban community area, and water sources. With an area of more than 821 meters above sea level, there are 6-102 patients in the area. Further, there is high population density in the area. From the comparison of model analysis results in Figure 8, the error point will be an area that has no risk or an area that is low risk. However, if considering from that information, then it must be a moderate risk or high-risk area.

Figure 12 shows model checking with data to there are any points on the map that the model has examined correctly or which ones are wrong in Random Tree from data 1000 points, correct 959 points and error 41 points. The point where the model analyzes has the most data errors. The area has agriculture and water sources with an area more than 821 meters above sea level. There are 6-102 patients in the area. Further, there is high population density in the area. From the comparison of model analysis results in Figure 9, the error point will be an area that has no risk or an area that is low risk. However, if considering from that information, then it must be a moderate risk or high-risk area.

Figure 13 shows model checking with data to any points on the map that the model has examined correctly or which ones are wrong in J48 from data 1000 points, correct 935 points and error 65 points. The point where the model analyzes has the most data errors. The area that has agriculture with an area of more than 821 meters is above sea level. There are 6-102 patients in the area. Further, there is high population density in the area. From the comparison of model analysis results in Figure 9, the error point will be an area that has no risk or an area that is low risk. However, if considering from that information, then it must be a moderate risk or high-risk area.

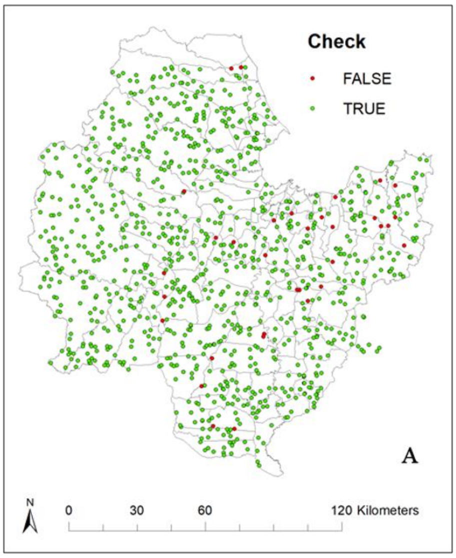

Fig. 11. Accuracy points from the Random forest model.

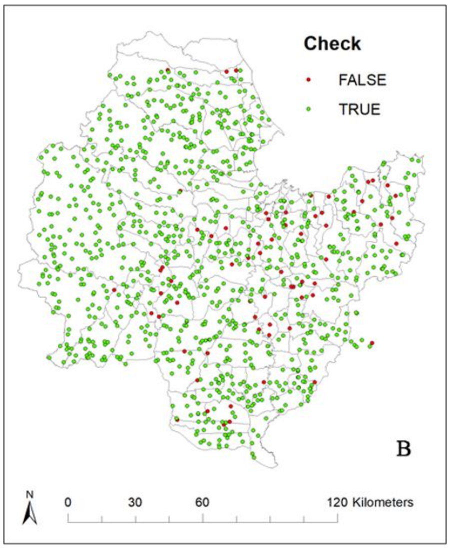

Fig. 12. Accuracy points from the Random Tree model.

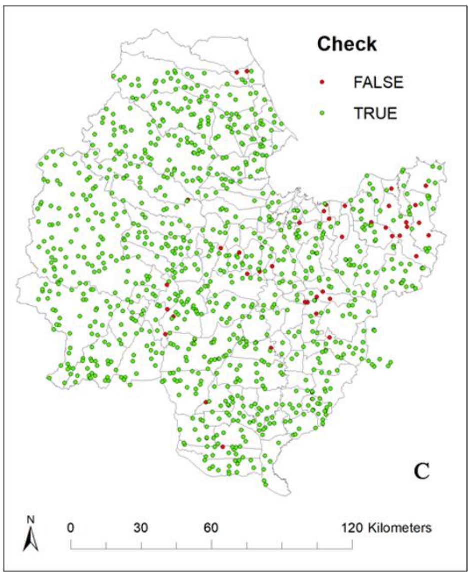

Fig. 13. Accuracy points from the Random Tree model.

Table 5 and Figure 14 show the comparison of different algorithms on the base of Accuracy, Root means square error (RMSE) and kappa statics. It can be seen that the data values for all 3 models are consistent and in the same direction as follows. The proposed algorithm has a significant accuracy difference compared to other algorithms. It has the maximum for Random Forest accuracy rate of 96.7%, which is the most accurate, while second is J48 accuracy rate of 95.9% and last is Random Tree with an accuracy rate of 93.5%.

When the accuracy is high, the value of the RMSE must be small because RMSE is a statistical measurement of volumes that are constantly changing. Calculations can be made to any series of values or any function that continuously fluctuates and used to compare the prediction accuracy of each model. The model which has the lowest RMSE is the best model [47]. Similarly, Kappa statistic is used to measure inter-rater reliability (and also Intra-rater reliability) for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of agreement occurring by chance.

Table 5. Performance comparison between different algorithms

Algorithms	Accuracy	RMSE	Kappa statistic
Random Forest	96.70	0.13	0.93
Random Tree	93.50	0.18	0.87
J48	95.90	0.14	0.92

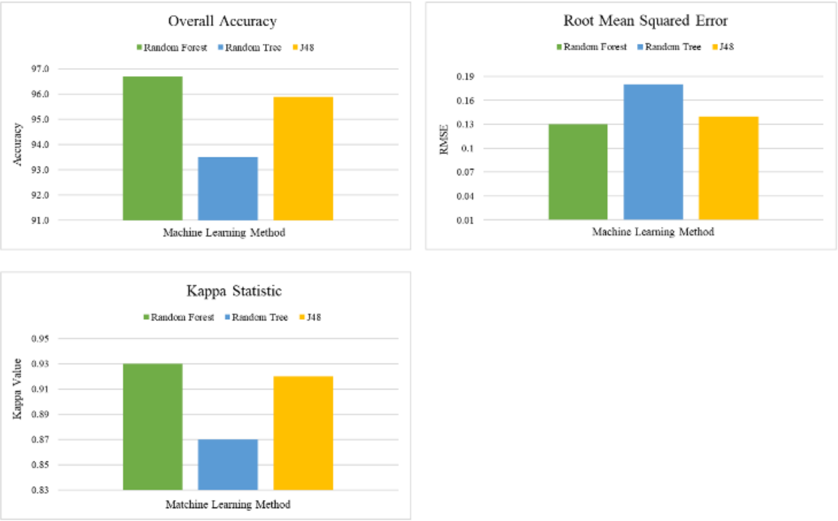

Fig. 14. A comparison of accuracy of each method.

Additionally, in this research, the accuracy of the model is evaluated when increasing the number of datasets used in learning and testing. It is found that when increasing the number of datasets, the accuracy will be higher. The error decreased. Random Forest algorithm is preferred, due to its consistent favorable performance. Accuracy assessment was performed on its RMSE and Accuracy measure. These values versus the number of datasets is plotted in Figure 15. It was evident from these graph that RMSE and Accuracy significantly improved up to approximately 2000th point. After that, it started to converge until approximately 4000th point, when no improvement was noticed. This result serves as a preliminary guideline on training the Random Forest model. Moreover, accuracy assessment on other parameter settings can follow the same suite, given a new set of areas, that may differ in terms of temperature, rainfall, and geospatial characteristics than the Surat Thani province, considered in this study. To Accuracy assessment of the model, 10 folds cross validation was used. The 10 folds cross validation results corresponded well with the above hypothesis: learning based on similar characteristic factors as the validating data results in better accuracy.

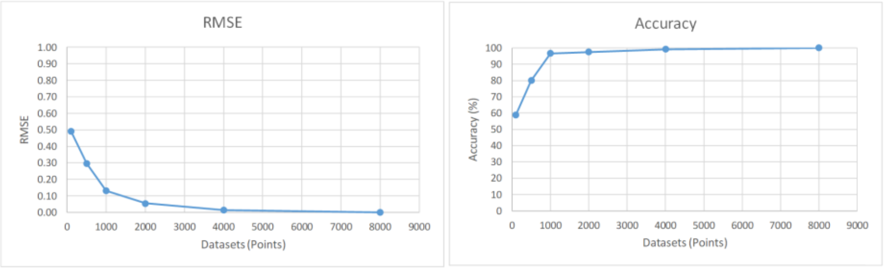

Fig. 15. RMSE and Accuracy versus the size of dataset.

4 Conclusion

From the analysis and the importance of factors affecting the outbreak of dengue fever in Surat Thani Province by using GIS and Data Mining techniques by using the data of each factor, the results are displayed in the form of geographic information data. It is able to determine which factors are suitable and at what level. When the data is analyzed with the Decision Tree and determined from the data of the results that show the map results, the most affective factors are LULC in the event that the area is an urban community area. This will result in a dengue epidemic outbreak due to the large number of people, which is consistent with the data for Population Density and the number of dengue patient's existent. These two factors are the second-most important factors. Next is the DEM; if the altitude is high compared to sea level, there is a greater risk of an outbreak. Finally, the level of rainfall is a factor. Since Surat Thani is a province with rain all year round, the amount of rainfall in every area is not very different (The first factors that were eliminated when analyzing data were temperature and humidity because the data was analyzed using data mining techniques. By reducing the value of these two factors, this did not result in the model's reliability changing, thus reducing these two factors). The models with the most accuracy are Random Forest (96.7%), J48 (95.9%), and Random Tree (93.5%). All three models do not have much difference and the accuracy of all 3 models is reliable because it is more than 90% accurate. We can either use these 3 models or choose the most accurate model, Random Forest, to support or consider when making a campaign planning decision to reduce the number of outbreaks of dengue fever in Surat Thani Province.

A suggestion from this research is that it may be presented by using different models in order to see the comparative errors more clearly. For future research, other factors should be studied in terms of the cases of dengue fever in an outbreak. For example, they may be used as a container factor, which will be detailed at the level of habitat, spawning, liking or attractiveness, causing mosquitoes to propagate at a place or container.

5 Acknowledgement

The authors are deeply grateful to the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani campus, Thailand. This research was financially supported by Prince of Songkla University, Surat Thani Campus.

6 References

[1] Somboonsak, P. (2020). Development Innovation to Predict Dengue Affected Area and Alert People with Smartphones. International Journal of Online and Biomedical Engineering (iJOE), 16(02), 62-79. https://doi.org/10.3991/ijoe.v16i02.12425

[2] Hafeez, S., Amin, M., & Munir, B. A. (2017). Spatial mapping of temporal risk to improve prevention measures: a case study of dengue epidemic in Lahore. Spatial and spatio-temporal epidemiology, 21, 77-85. https://doi.org/10.1016/j.sste.2017.04.001

[3] Wongkoon, S., Jaroensutasinee, M., & Jaroensutasinee, K. (2016). Spatio-temporal climate-based model of dengue infection in Southern, Thailand. Tropical Biomedicine, 33(1), 55-70. https://doi.org/10.1016/s1995-7645(12)60034-0

[4] Asmahani, A., MR, M. N., & Harsuzilawati, M. (2010). Spatial mapping of dengue incidence: A case study in hulu langat district, Selangor, Malaysia. International Journal of Geological and Environmental Engineering, 4(7), 251-255.

[5] Sermkarndee P., Manwicha, J., and Khunchoo R. (2016). Risk Areas Analysis of Dengue Fever Using Geographic Information Systems, Hatyai District, Songkhla Province. Hat Yai National and International Academic Conference, 7, 1355-1365.

[6] Zambrano, L. I., Sierra, M., Lara, B., Rodríguez-Núñez, I., Medina, M. T., Lozada-Riascos, C. O., & Rodríguez-Morales, A. J. (2017). Estimating and mapping the incidence of dengue and chikungunya in Honduras during 2015 using Geographic Information Systems (GIS). Journal of infection and public health, 10(4), 446-456. https://doi.org/10.1016/j.jiph.2016.08.003

[7] Mutheneni, S. R., Mopuri, R., Naish, S., Gunti, D., & Upadhyayula, S. M. (2018). Spatial distribution and cluster analysis of dengue using self organiself-organizingdhra Pradesh, India, 2011–2013. Parasite epidemiology and control, 3(1), 52-61. https://doi.org/10.1016/j.parepi.2016.11.001

[8] Santos, C. A. G., Guerra-Gomes, I. C., Gois, B. M., Peixoto, R. F., Keesen, T. S. L., & da Silva, R. M. (2019). Correlation of dengue incidence and rainfall occurrence using wavelet transform for Joao Pessoa city. Science of The Total Environment, 647, 794-805. https://doi.org/10.1016/j.scitotenv.2018.08.019

[9] Her, Z., Kam, Y. W., Gan, V. C., Lee, B., Thein, T. L., Tan, J. J., & Renia, L. (2017). Severity of plasma leakage is associated with high levels of interferon gamma-inducible protein 10, hepatocyte growth factor, matrix metalloproteinase 2 (MMP-2), and MMP-9 during dengue virus infection. Journal of Infectious Diseases, 215(1), 42-51. https://doi.org/10.1093/infdis/jiw494

[10] Bravo, L., Roque, V. G., Brett, J., Dizon, R., & L'Azou, M. (2014). Epidemiology of dengue disease in the Philippines (2000–2011): a systematic literature review. PLoS neglected tropical diseases, 8(11). https://doi.org/10.1371/journal.pntd.0003027

[11] Xu, Z., Bambrick, H., Yakob, L., Devine, G., Lu, J., Frentiu, F. D., & Hu, W. (2019). Spatiotemporal patterns and climatic drivers of severe dengue in Thailand. Science of The Total Environment, 656, 889-901. https://doi.org/10.1016/j.scitotenv.2018.11.395

[12] Department of Disease Control, 2019. Forecast report Dengue fever of 2019. Available from: https://ddc.moph.go.th/uploads/ckeditor/6f4922f45568161a8cdf4ad2299f6d23/files/Dangue/Prophecy/2562.pdf and https://ddc.moph.go.th/uploads/ckeditor/6f4922f45568161a8cdf4ad2299f6d23/files/Dangue/Situation/2562/DHF%20%201.pdf. Accessed: March 2019.

[13] Fuller, D. O., Troyo, A., Alimi, T. O., & Beier, J. C. (2014). Participatory risk mapping of malaria vector exposure in northern South America using environmental and population data. Applied Geography, 48, 1-7. https://doi.org/10.1016/j.apgeog.2014.01.002

[14] Stevens, K. B., & Pfeiffer, D. U. (2011). Spatial modelling of disease using data-and knowledge-driven approaches. Spatial and spatio-temporal epidemiology, 2(3), 125-133. https://doi.org/10.1016/j.sste.2011.07.007

[15] Sammatat, S., Boonsith, N., & Lekdee, K. (2014). Spatial mathematical analysis: an application to mapping of dengue hemorrhagic fever in Thailand.

[16] Agarwal, N., Koti, S. R., Saran, S., & Kumar, A. S. (2018). Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi,

India. Current Science, 114(11), 2281-2291. https://doi.org/10.18520/cs/v114/i11/2281-2291

[17] Attaway, D. F., Jacobsen, K. H., Falconer, A., Manca, G., & Waters, N. M. (2016). Risk analysis for dengue suitability in Africa using the ArcGIS predictive analysis tools (PA tools). Acta tropica, 158, 248-257. https://doi.org/10.1016/j.actatropica.2016.02.018

[18] Minale, A. S., & Alemu, K. (2018). Mapping malaria risk using geographic information systems and remote sensing: The case of Bahir Dar City, Ethiopia. Geospatial health, 13(1). https://doi.org/10.4081/gh.2018.660

[19] Somard, J., Suwanlee, R. S., Turnbull, N., & Phommat, T., (2017). Analyzing dengue fever risk areas using geographic information systems in Dome Pradis Sub-district, Nam Yuen District, Ubon Ratchathani Province. J Med Health Sci. 24(3), 65-76.

[20] Preechapanich, O., & Thernmontri, S. (2015). A Geographic Information System for Supporting the Surveillance of Dengue Infection in Songkhla Province. Thaksin University Journal, 18(3), 161-169.

[21] The Office of Disease Prevention and Control 3 ChonBuri. (2012). Prediction risky area of dengue haemorrhagic fever among eastern-seaboard part of Thailand 2012. Available from: www.interfetpthailand.net/forecast/_files/report2012/report_2012_11_no12.pdf. Accessed: March 2019.

[22] Thai Meteorological Department. (2018). Rainfall data, temperature and humidity. Available from: https://www.tmd.go.th. Accessed: March 2019.

[23] National Statistical Office Thailand. (2019). Provincial Population Density data. Available from: http://www.nso.go.th/ sites/2014/Pages/home.aspx. Accessed: March 2019.

[24] Surat Thani Provincial Health Office. (2019). Weekly dengue data. Available from: http://www.stpho.go.th/. Accessed: March 2019.

[25] Yue, Y., Sun, J., Liu, X., Ren, D., Liu, Q., Xiao, X., & Lu, L. (2018). Spatial analysis of dengue fever and exploration of its environmental and socio-economic risk factors using ordinary least squares: A case study in five districts of Guangzhou City, China, 2014. International Journal of Infectious Diseases, 75, 39-48. https://doi.org/10.1016/j.ijid.2018.07.023

[26] Liu, K., Zhu, Y., Xia, Y., Zhang, Y., Huang, X., Huang, J., & Hu, W. (2018). Dynamic spatiotemporal analysis of indigenous dengue fever at street-level in Guangzhou city, China. PLoS neglected tropical diseases, 12(3), e0006318. https://doi.org/10.1371/journal.pntd.0006318

[27] Chafiq, T., Ouadoud, M., Oulidi, H. J., & Fekri, A. (2018). Application of Data Integrity Algorithm for Geotechnical Data Quality Management. International Journal of Interactive Mobile Technologies, 12(8), 85-95. https://doi.org/10.3991/ijim.v12i8.9569

[28] Lam, K. C., Bryant, R. G., & Wainright, J. (2015). Application of spatial interpolation method for estimating the spatial variability of rainfall in semiarid New Mexico, USA. Mediterranean Journal of Social Sciences, 6(4), 108-108. https://doi.org/10.5901/mjss.2015.v6n4s3p108

[29] Yang, M. (2015). Benchmarking rainfall interpolation over the Netherlands. University of Twente Faculty of Geo-Information and Earth Observation (ITC).

[30] Umor, S. M., Mokhtar, M., Surip, N., & Ahmad, A. (2007). Generating a dengue risk map (DRM) based on environmental factors using remote sensing and GIS technologies. In 28th Asian Conference on Remote Sensing 2007, ACRS 2007 (pp. 867-881).

[31] Puttinaovarat, S., & Horkaew, P. (2019). Application Programming Interface for Flood Forecasting from Geospatial Big Data and Crowdsourcing Data. International Journal of Interactive Mobile Technologies, 13(11). https://doi.org/10.3991/ijim.v13i11.11237

[32] Puttinaovarat, S., & Horkaew, P. (2020). Flood Forecasting System Based on Integrated Big and Crowdsource Data by Using Machine Learning Techniques. IEEE Access, 8, 5885-5905. https://doi.org/10.1109/access.2019.2963819

[33] Ong, J., Liu, X., Rajarethinam, J., Kok, S. Y., Liang, S., Tang, C. S., & Yap, G. (2018). Mapping dengue risk in Singapore using Random Forest. PLoS neglected tropical diseases, 12(6), e0006587. https://doi.org/10.1371/journal.pntd.0006587

[34] Thiptida, V. (2013). Lease Approval Using Data Mining Techniques. Department of Computer and Communication Technology Faculty of Engineering, Dhurakij Pundit University.

[35] Czajkowski, M., & Kretowski, M. (2019). Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. Expert Systems with Applications, 137, 392-404. https://doi.org/10.1016/j.eswa.2019.07.019

[36] Sáez, J. A., Luengo, J., & Herrera, F. (2016). Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure. Neurocomputing, 176, 26-35. https://doi.org/10.1016/j.neucom.2014.11.086

[37] Loh, W. Y. (2014). Fifty years of classification and regression trees. International Statistical Review, 82(3), 329-348. https://doi.org/10.1111/insr.12016

[38] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2019.05.028

[39] Speiser J.L., Durkalski V.L., and Lee W.M. 2015. Random forest classification of etiologies for an orphan disease. Statistics in Medicine. Volume 34 (5): 887-899. https://doi.org/10.1002/sim.6351

[40] Speiser, J. L., Durkalski, V. L., & Lee, W. M. (2015). Random forest classification of etiologies for an orphan disease. Statistics in medicine, 34(5), 887-899. https://doi.org/10.1002/sim.6351

[41] Mao, W., & Wang, F. (2012). Chapter 8 - Cultural Modeling for Behavior Analysis and Prediction. New Advances in Intelligence and Security Informatics, 91-102. https://doi.org/10.1016/b978-0-12-397200-2.00008-7

[42] Louppe, Gilles. (2014). Understanding Random Forests: From Theory to Practice. University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science.

[43] Mishra, A. K., & Ratha, B. K. (2016). Study of random tree and random forest data mining algorithms for microarray data analysis. International Journal on Advanced Electrical and Computer Engineering, 3(4), 5-7.

[44] Esri. (2019). Train Random Trees Classifier. Available from: https://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/train-random-trees-classifier.htm. Accessed: August 2019.

[45] Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22). https://doi.org/10.5120/17314-7433

[46] Shafie, F. A., Tahir, M. P. M., & Sabri, N. M. (2012). Aedes mosquito's resistance in urban community setting. Procedia-Social and Behavioral Sciences, 36, 70-76. https://doi.org/10.1016/j.sbspro.2012.03.008

[47] Chayanin, B., & Nat, K. (2017). A Comparative Prediction Accuracy of Hybrid Time Series Models. Science and Technology, 25(2), 177-190.

7 Authors

Benjawan Hnusuwan received the B.Sc. degree in information technology from the Prince of Songkla University, Thailand, in 2017, where she is currently pursuing the M.Sc. degree in applied mathematics and computing science. Her research interest includes the Dengue Hemorrhagic Fever and Information Technology.

Siriwan Kajornkasirat is an assistant professor at Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus. Her research interest includes Dengue Hemorrhagic Fever, Data science and IoT.

Supattra Puttinaovarat is an assistant professor at Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus. Her research interest includes Geographic Information System, Remote Sensing, Machine Learning and Information Technology.

Article submitted 2020-06-21. Resubmitted 2020-07-10. Final acceptance 2020-07-11. Final version published as submitted by the authors.