



Leptospirosis modelling using hydrometeorological indices and random forest machine learning

Veianthan Jayaramu¹ · Zed Zulkafli¹ · Simon De Stercke² · Wouter Buytaert² · Fariq Rahmat³ · Ribhan Zafira Abdul Rahman³ · Asnor Juraiza Ishak³ · Wardah Tahir⁴ · Jamalludin Ab Rahman⁵ · Nik Mohd Hafiz Mohd Fuzi⁶

Received: 30 May 2022 / Revised: 21 December 2022 / Accepted: 26 December 2022 / Published online: 31 January 2023
© The Author(s) under exclusive licence to International Society of Biometeorology 2023

Abstract

Leptospirosis is a zoonosis that has been linked to hydrometeorological variability. Hydrometeorological averages and extremes have been used before as drivers in the statistical prediction of disease. However, their importance and predictive capacity are still little known. In this study, the use of a random forest classifier was explored to analyze the relative importance of hydrometeorological indices in developing the leptospirosis model and to evaluate the performance of models based on the type of indices used, using case data from three districts in Kelantan, Malaysia, that experience annual monsoonal rainfall and flooding. First, hydrometeorological data including rainfall, streamflow, water level, relative humidity, and temperature were transformed into 164 weekly average and extreme indices in accordance with the Expert Team on Climate Change Detection and Indices (ETCCDI). Then, weekly case occurrences were classified into binary classes “high” and “low” based on an average threshold. Seventeen models based on “average,” “extreme,” and “mixed” indices were trained by optimizing the feature subsets based on the model computed mean decrease Gini (MDG) scores. The variable importance was assessed through cross-correlation analysis and the MDG score. The average and extreme models showed similar prediction accuracy ranges (61.5–76.1% and 72.3–77.0%) while the mixed models showed an improvement (71.7–82.6% prediction accuracy). An extreme model was the most sensitive while an average model was the most specific. The time lag associated with the driving indices agreed with the seasonality of the monsoon. The rainfall variable (extreme) was the most important in classifying the leptospirosis occurrence while streamflow was the least important despite showing higher correlations with leptospirosis.

Keywords Leptospirosis · Hydrometeorological indices · Cross-correlation analysis · Random forest · Variable importance · Feature selection

✉ Zed Zulkafli
zeddiyana@upm.edu.my; zed.zulkafli@gmail.com

- ¹ Department of Civil Engineering, Universiti Putra Malaysia, Serdang, Malaysia
- ² Department of Civil and Environmental Engineering, Imperial College London, London, UK
- ³ Department of Electrical and Electronic Engineering, Universiti Putra Malaysia, Serdang, Malaysia
- ⁴ Flood Control Research Group, Faculty of Civil Engineering, Universiti Teknologi Mara, Shah Alam, Malaysia
- ⁵ Department of Community Medicine, Kulliyah of Medicine, International Islamic University Malaysia, Kuantan, Malaysia
- ⁶ Kelantan State Health Department, Ministry of Health Malaysia, Kota Bharu, Kelantan, Malaysia

Introduction

Leptospirosis is a zoonotic disease caused by pathogenic spirochetes of the genus *Leptospira* (Picardeau 2013). The bacteria infect humans directly through contact with the urine of an infected host or indirectly through contact with a contaminated environment, entering via an injured skin and/or mucous membrane (Ansdell 2017). The disease exists in both temperate and tropical countries, although higher incidence rates are reported in the latter (10–100 compared to 0.1 to 1 per 100,000 population per year, WHO 2003). The higher incidence rates in tropical regions could be due to their year-round warm and humid climate, which is favorable to the survivability and dynamics of leptospires (Adler and de la Peña

Moctezuma 2010). Furthermore, heavy rainfall and consequent flooding events have contributed to more widespread infections (Barcellos & Sabroza 2001; Mohd Radi et al. 2018; Togami et al. 2018; Sehgal et al. 2002; Ding et al. 2019). Climate change and the associated increase in the frequency of hydrometeorological extreme events are expected to further aggravate the risk of infection (Lau et al. 2010; Picardeau 2013).

To better understand the mechanism behind leptospirosis transmission, several different approaches in its statistical modelling have been explored. While many have considered the spatial dependency (Lau et al. 2012; Schneider et al. 2012; Suwanpakdee et al. 2015; Vega-Corredor & Opadeyi 2014; Zhao et al. 2016; Mayfield et al. 2018; Mohammadinia et al. 2017; Sánchez-Montes et al. 2015), fewer studies have analyzed the drivers behind the occurrence, transmission, and outbreak using the time series (Chadsuthi et al. 2012; Desvars et al. 2011; Joshi et al. 2017; Weinberger et al. 2014). Additionally, most have employed conventional statistical modelling techniques, which inadequately handled the nonlinearity between leptospirosis and its risk factors (Dhewantara et al. 2019). The complex mechanism of leptospirosis transmission due to the involvement of multiple variables impedes the models' ability of explaining the disease trends (WHO 2011). Machine learning models, in contrast, can capture more complex patterns and therefore predict the output with higher accuracy (Carvajal et al. 2018; Guo et al. 2017; Hu et al. 2018; Ahangarcani et al. 2019). However, they are often treated as black boxes when the objective is to optimize predictive performance as opposed to gaining process insights.

Nevertheless, knowledge extraction is possible with the use of interpretable machine learning algorithms. Random forest machine learning (Breiman 2001, overview in the subsection “[Structure and algorithm](#)”) is one that allows insight into feature importance. Unlike the neural network and support vector machine, the random forest algorithm uses tree-based decision-making and can rank the features involved during the model training based on how well they contribute to the classification of output classes. It has been applied for predicting water-borne and vector-borne diseases, such as cholera (Campbell et al. 2020), dengue (Carvajal et al. 2018; Khan et al. 2017; Zhao et al. 2020), malaria (Barradas-Bautista 2020), and tick-borne encephalitis (Uusitalo et al. 2020). It has also been used to model animal leptospirosis based on annual precipitation and temperature as well as socio-economic and landscape factors (Zakharova et al. 2021). In Zakharova et al. (2021), independent variables were ranked based on the important metric of Gini that reflects the variable's responsibility in splitting the output. However, the study did not further optimize the model by removing the less important (lower ranked)

variables. Eliminating less important or irrelevant variables can reduce the complexity of the model, which improves its run time, comprehensibility, and performance (Kumar and Minz 2014).

In this study, cross-correlation analysis was used, and the capabilities of the random forest algorithm were leveraged to answer the following research questions:

- (1) What hydrometeorological indices are highly cross-correlated with leptospirosis and important in classifying the disease occurrence?
- (2) Does prediction capacity change according to the type of index used as model features, whether in the form of average or extreme indices or their combination?

Hydrometeorological variability in the form of averages and extremes has been used as drivers in past modelling studies. For example, simple average and extreme hydrometeorological indices, i.e., mean, sum, minimum, and maximum, have been investigated (Chadsuthi et al. 2012; Cunha et al. 2022; Desvars et al. 2011; Gómez et al. 2022; Kupek et al. 2000; Mohd Radi et al. 2018; Rahmat et al. 2019; Schneider et al. 2012; Sumi et al. 2017; Weinberger et al. 2014), while more elaborate covariates that represented extreme dry and wet conditions have also been used (Tassinari et al. 2008; Sánchez-Montes et al. 2015; Rahayu et al. 2018; Ding et al. 2019; Ehelepola et al. 2019). However, none of the studies have systematically analyzed and compared the effects of different variables and their average and extreme indices on case predictions. In this study, 17 random forest classification models of leptospirosis occurrence were implemented to identify the predictive performance of different indices (“average,” “extreme,” and “mixed” indices, defined in “[Model configuration based on classes of indices](#)”) used. Additionally, variable importance was assessed based on an analysis of correlation that considered the lag, as well as the mean decrease Gini (MDG, defined in “[Feature subset selection](#)”) score from the random forest model.

Materials and methods

Study area

The case study is in Kelantan, a northeastern state of Peninsular Malaysia that experiences extreme monsoonal rainfall annually, often leading to extended periods of floods (Ismail and Haghroosta 2018). The state consists of 10 districts bordering Thailand in the northwest, Perak in the southwest, Pahang in the south, and Terengganu in the southeast. The state recorded the highest incidence rate in 2015 following a massive flooding event in December 2014. The study area is three flood-prone (Department of Irrigation and Drainage

Malaysia [DID] 2017) Kelantan districts with the highest incidence rates (per 100,000 population) for the 10-year period from 2011 to 2020, i.e., Pasir Mas (638.0), Tumpat (344.5), and Kota Bharu (310.6). The location of the study area is depicted in Fig. 1.

Data collection and processing

Leptospirosis case data

Leptospirosis is considered endemic in Malaysia with cases occurring throughout the year (Benacer et al. 2016). The weekly number of probable and lab-confirmed leptospirosis cases from January 2011 to November 2020 was retrieved upon request from the Kelantan State Department of Health. In total, 6895 cases were reported throughout the period for the entire state. Nearly half of the total number of cases came from the selected districts, i.e., Pasir Mas, Tumpat, and Kota Bharu. These three districts were grouped into one model which resulted in 517 weekly records of case numbers. Each week contains the total number of cases aggregated over the

three districts. Since the number of cases for these districts was used in a single lumped model, the spatial variability in the study area was not explicitly considered.

Hydrometeorological data

Five types of hydrometeorological data were used in this study, i.e., rainfall, streamflow, water level, relative humidity, and temperature. They are daily data which span the 10-year period between 01/01/2011 and 31/12/2020. Rainfall, streamflow, and water level data were obtained upon request from the Department of Irrigation and Drainage Malaysia (DID) who is responsible for hydrological monitoring, while temperature and relative humidity data were available with purchase from the Malaysian Meteorological Department (MetMalaysia), who monitors the weather conditions of the country. For temperature, three levels of data were collected including daily minimum (M), mean (M), and maximum (X) temperature. MetMalaysia also provides rainfall data from the weather stations.

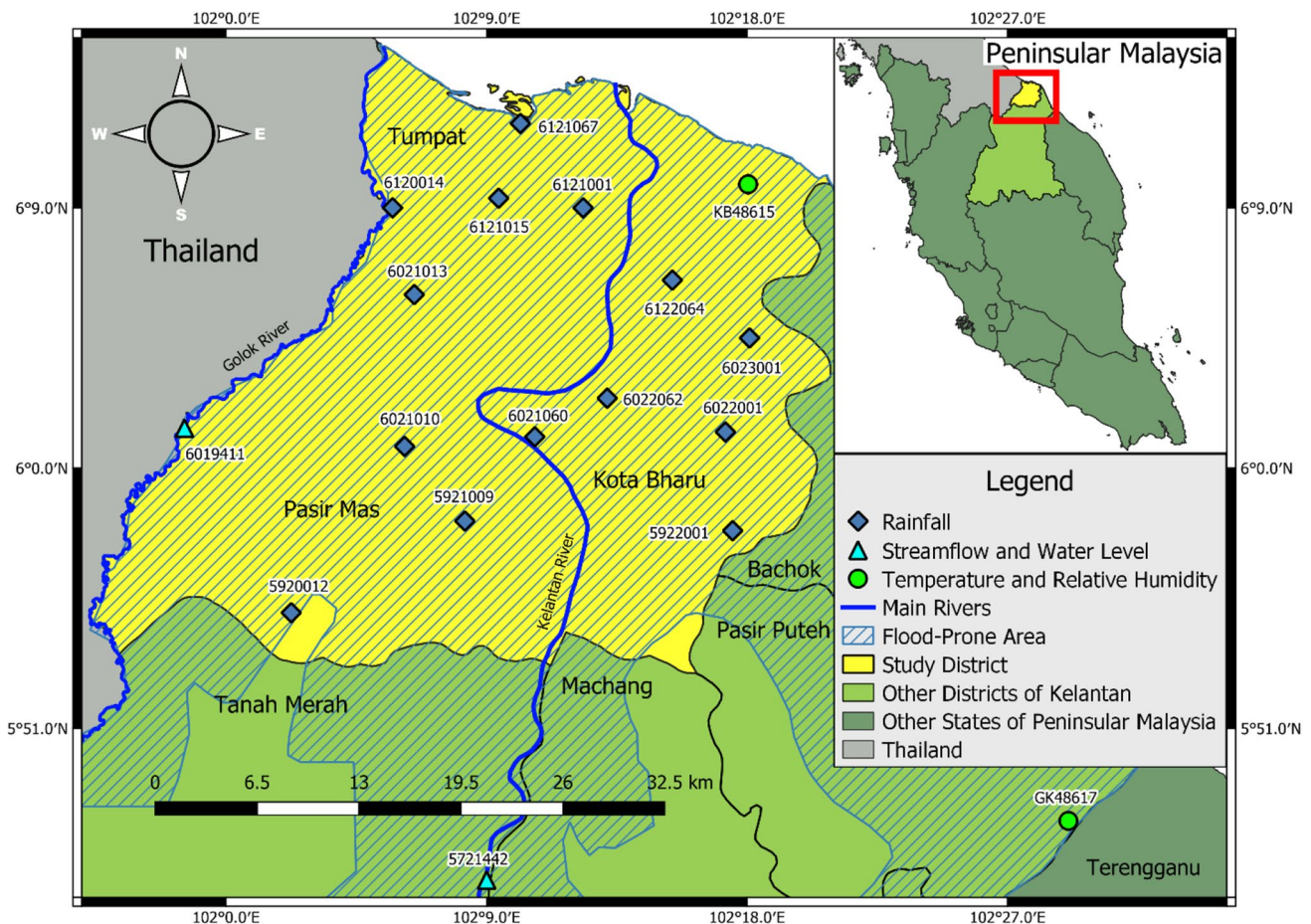


Fig. 1 The location of the study area, hydrometeorological stations, and flood-prone areas

Hydrometeorological index calculation

The study adopts hydrometeorological indices prescribed by the Expert Team on Climate Change Detection and Indices (ETCCDI), a program that was working on characterizing the climate variability and change (Peterson et al. 2001). These indices are derived from daily data to represent the frequency, intensity, and duration of hydrometeorological events. Although the ETCCDI has been recently discontinued, the indices are still valid and usable. The use of externally defined thresholds, with a few exceptions, was particularly intended to reduce subjectivity. In this study, average and extreme indices were produced, each of which with three subsets of indices, i.e., simple, fixed, and relative (Zhang et al. 2011):

- i. A simple extreme index was calculated based on the maximum and minimum values of the week.
- ii. A fixed extreme index is the number of days of the week with values exceeding beyond a specified limit. The specified limits are those defined by ETCCDI, DID, and local studies.
- iii. A relative extreme index is the number of days of the week with values exceeding an extreme percentile.
- iv. A simple average index was calculated based on the mean value of the week.
- v. A fixed average index is the number of days of the week with values between the same extreme limits used for deriving the fixed extreme indices.
- vi. A relative average index is the number of days of the week with values between the same extreme percentiles used for deriving the relative extreme indices.

Tables 1 and 2 summarize the indices calculated for the hydrometeorological variables of this study. The rainfall and temperature indices followed the naming convention used by ETCCDI. The indices of relative humidity, streamflow, and water level were named similarly for consistency.

Average indices that are not defined by the ETCCDI were named similarly as the extreme indices. ETCCDI prescribes the relative and fixed thresholds for rainfall and temperature but not for streamflow, water level, and relative humidity. The derivation of indices was adapted according to the following:

The relative thresholds for streamflow, water level, and relative humidity were selected using trial-and-error based on the percentile value that resulted in the highest correlation of the index to leptospirosis cases. This only involved the percentile values that were used by ETCCDI for rainfall and temperature, which include the 90th, 95th, and 99th percentiles.

The fixed threshold for streamflow was determined by averaging the streamflow values during a known flooding event.

The fixed threshold for water level was taken from the danger limit provided by the DID.

The fixed thresholds of temperature defined by ETCCDI could not be applied to the study due to different climates. Thus, the thresholds were taken from the local studies (Jamaludin et al. 2015).

Relative humidity was not aggregated into fixed average and extreme days since there was no literature found to support a suitable threshold for this study.

The naming structure of indices starts with the abbreviation of hydrometeorological data followed by the index used and ends with the station’s ID that is separated by an underscore. For example, the weekly mean (index) rainfall (RF) from 6121067 station was named as “RFmean_6121067.”

Cross-correlation analysis for lag identification

Cross-correlation (Equation S1) was analyzed between the weekly aggregated hydrometeorological indices and leptospirosis cases at multiple temporal lags. This is to select the

Table 1 The index ID of weekly aggregated average and extreme hydrometeorological indices for each class of index. For a full explanation of the annotation, refer to Table 2

Condition	Index class	Variable				
		Rainfall	Temperature	Humidity	Streamflow	Water level
Average	Simple	Mean	Mean	Mean	Mean	Mean
	Fixed	d1mm.d50mm	d23.d34		dnotFlood	dnotDanger
	Relative	d5p.d95p	d10p.d90p	d10p.d90p	d1p.d99p	d1p.d99p
Extreme	Simple	Maximum	Maximum	Maximum	Maximum	Maximum
			Minimum	Minimum		
	Fixed	d50mm	d34		dFlood	dDanger
		d23				
	Relative	d95p	d90p	d90p	d99p	d99p
			d10p	d10p		

Table 2 The description of each index ID of weekly aggregated average and extreme hydrometeorological indices

Index ID	Description
Average	
Mean	Mean rainfall, streamflow, water level, relative humidity, or temperature of the week
d1mm.d50mm	The number of days of the week with an average rainfall between 1 and 50 mm
d23.d34	The number of days of the week with an average temperature between 23 and 34 °C
dnotFlood	The number of days of the week with an average streamflow of not exceeding the flooding threshold
dnotDanger	The number of days of the week with an average water level of not exceeding the danger limit
d5p.d95p	The number of days of the week with an average rainfall between the 5th and 95th percentiles of the base period
d10p.d90p	The number of days of the week with an average relative humidity or temperature between the 10th and 90th percentiles of the base period
d1p.d99p	The number of days of the week with an average streamflow or water level between the 1st and 99th percentiles of the base period
Extreme	
Maximum	Maximum rainfall, streamflow, water level, humidity, or temperature of the week
Minimum	Minimum relative humidity or temperature of the week
d50mm	The number of days of the week with an extreme rainfall exceeding 50 mm
d34	The number of days of the week with an extreme maximum temperature exceeding 34 °C
d23	The number of days of the week with an extreme minimum temperature not exceeding 23 °C
dFlood	The number of days of the week with an extreme streamflow exceeding the flooding threshold
dDanger	The number of days of the week with an extreme water level exceeding the danger limit
d95p	The number of days of the week with an extreme rainfall exceeding the 95th percentile of the base period
d90p	The number of days of the week with an extreme relative humidity or temperature exceeding the 90th percentile of the base period
d10p	The number of days of the week with an extreme relative humidity or temperature not exceeding the 10th percentile of the base period
d99p	The number of days of the week with an extreme streamflow or water level exceeding the 99th percentile of the base period

lag at which the correlation is highest and adjust the indices time series accordingly in the model setup, as lag has been demonstrated to be an important factor in the hydrometeorological controls of leptospirosis (Rahmat et al. 2020). The maximum lag allowed was 52 weeks to observe the correlation patterns of up to a year between leptospirosis and the lagged hydrometeorological indices. This resulted in 465 weeks of complete case (lagged) and indices records. A detailed explanation of the cross-correlation analysis can be found in Supporting Information.

Binary classification of leptospirosis

A threshold was selected to classify the number of leptospirosis cases into high and low, based on the average weekly cases. The average number of weekly cases over the period of the study was six cases, and the classification resulted in 171 (37%) weeks with high cases (more than six) and 294 (63%) weeks with low cases (six and below).

Model configuration based on classes of indices

There were 17 configurations of models in total based on classes of indices (see Fig. 3). Filter method (Kumar and

Minz 2014) was used to select the features for models A5 and E5. This method selects the features based on the highest correlation to the dependent variable without involving the model's learning algorithm. The random forest models were developed using the *caret* package, short for classification and regression training (Kuhn 2008), within the R computing environment. The normalized datasets of final models and R code for training and testing the models are available on GitHub (link: https://github.com/VeianthanJayaramu/Kelantan_leptospirosis_modelling).

Random forest classification model development

Structure and algorithm

Random forest (Breiman 2001) is an ensemble of trees grown from the bagging method that derives resampled datasets with records duplicated from the original training set. In each tree, the nodes containing the response variable are split recursively by the selected features (from random subsets) until they meet a specified node size. The randomness in the feature subsets decorrelates the trees, which reduces the algorithm's sensitivity to multicollinearity effects. Finally, the outputs are aggregated across the trees

to finalize them based on majority voting. The splitting rule used in this study was based on the Gini impurity, which measured the probability of misclassification as a result of the splitting of features at the nodes. The Gini impurity of a node (G_n) was calculated as one minus the summation of the squared probabilities of the classified outputs, high (O_h) and low (O_l). More specifically (Eq. 1):

$$G_n = 1 - \left(\frac{O_h}{O_h + O_l} \right)^2 - \left(\frac{O_l}{O_h + O_l} \right)^2 \quad (1)$$

The total Gini impurity (G_{sn}) for the following right (G_r) and left (G_l) sub-nodes is the probability weighted from the fractions (p) of data sent by the splitting node (Eq. 2).

$$G_{sn} = p_r G_r + p_l G_l \quad (2)$$

Tuning hyperparameters

mtry, *nree*, and *nodesize* are the hyperparameters that govern the growth and density of trees in the forest. The *mtry* parameter controls the number of randomly selected candidate predictors at each splitting node, while the *nree* parameter determines the number of trees the forest should have. The *nodesize* parameter limits the depth of individual trees with a specified minimum number of observations for the terminal node. The optimal set of hyperparameters was determined using the grid search method that ran the models under all the possible combinations of the tuning parameters. *mtry* parameter was set to range from one to the total number of features present in the model input datasets while *nree* was set to range between 200 and 700 trees at the interval of 50 trees. The *nodesize* was searched from one to 10 minimum records in the terminal node.

Data partitioning

The model input data was split into 80% for the training set and 20% for the testing set, which is an acceptable split per Ucar et al. (2020). A higher training ratio was used to obtain a more reliable relative importance measure that gets calculated during model training. The split was done in such a way that preserved the proportion of binary case classes in both training and testing sets. The training set contains 137 (37%) weeks with a high number of cases and 236 (63%) weeks with low cases. On the other hand, the testing set contains 34 (37%) weeks with high cases and 58 (63%) weeks with a low number of cases. Additionally, a 10 k-fold cross-validation was conducted for all the models to prevent overfitting (Santos et al. 2018). It is an internal model fitting procedure within the training set that was carried out a total of 10 times, each of which on a sub-training set that consisted of 90% of the

total training data selected at random. The remaining 10% was then used for the validation. The performance metrics reported are thus the average results of 10 training and testing processes. This 10 k-fold cross-validation procedure was repeated thrice to reduce noise in the estimation of model performance.

Feature subset selection

The random forest algorithm computed a score called mean decrease Gini (MDG) during the model training, which indicated the importance of each predictor to the model. The MDG measured the total decrease (ΔG , Eq. 3) in Gini score before (G_n) and after (G_{sn}) the node split, which was then averaged across the trees (Zhang et al. 2019). The higher the MDG score, the more important the variable was to the model since it contributed more towards the good classifications of high and low outputs. Therefore, this feature is considered more important compared to other features with lower MDG values. The preliminary models developed based on the configurations in “[Model configuration based on classes of indices](#)” were optimized by retaining the highly important hydrometeorological indices. Essentially, the indices of the preliminary models were ranked based on the MDG scores. Then, to obtain an optimal subset of the training data, the models underwent a sequential forward selection that consecutively added the ranked features to the subset one by one, starting from the most important variable. The training accuracy of each subset was evaluated, and when it started to deteriorate, the feature space was selected for the final model development. Some of the models’ accuracy fluctuated with large differences over the increasing size of subsets. In this case, several subsets with relatively higher accuracy were selected and used to train the final models. Among the models, the subset that exhibited the highest training accuracy was selected. The selected subset was then tuned for the best model parameters and tested with the 20% holdout set to calculate the prediction accuracy.

$$\Delta G = G_n - G_{sn} \quad (3)$$

Model performance measurement

The models were evaluated using the testing set accuracy, sensitivity, and specificity. Accuracy is the percentage of correct classification of both high and low cases; sensitivity is the percentage of predicted high cases over the actual high cases; and specificity is the percentage of predicted low cases over the actual low cases (Van Stralen et al. 2009; Glaros and Kline 1988). The formulae used to

calculate these metrics are presented in Table S1 of Supporting Information.

Receiver operating characteristic (ROC) analysis

ROC analysis was conducted to produce a curve that aids in finding the optimal operating points for best separating the outputs into a high and low number of case predictions. Each operating point that varied between 0 and 1 separated the binary outputs based on their probability values. The sensitivity and specificity of the outputs were calculated at each operating point, and a curve of sensitivity against 1-specificity was plotted. The classifier that gives the curve closer to the top-left corner (maximized sensitivity and specificity) indicates a better performance based on accuracy in classifying the high and low cases of leptospirosis.

Results

Cross-correlation analysis (CCA)

A total of 164 hydrometeorological indices were derived to develop 17 models. Of these, 155 predictors were weakly correlated, and one predictor (SFmean_5721442) was moderately correlated with leptospirosis cases. Eight predictors were considered negligible according to Schober et al. (2018) since their correlations were below 0.1. However, out of eight, only two hydrometeorological indices, i.e., TMd23.d34_KB48615 and TMd23.d34_GK48617, were not included in the models since their correlations were insignificant ($p > 0.05$). The result of cross-correlation analyses is presented in Supplementary Information Figures S1–S5.

The rainfall indices were positively correlated with leptospirosis cases at shorter lags of up to 15 weeks and negatively correlated at longer lags especially at the 35 weeks lag and higher. The highest positive correlations between the d50mm and d95p of rainfall (i.e., the number of days of extreme rainfall exceeding the 50 mm and 95th percentile thresholds) and leptospirosis cases were at an earlier lag of 3 weeks. The

highest positive correlations between the d1mm.d50mm and d5p.d95p of rainfall (i.e., the number of days of average rainfall between the lower and upper extreme limits) and leptospirosis cases were instead at a later lag of 9 weeks.

Meanwhile, the streamflow indices were correlated with leptospirosis cases up to 16 weeks lag. The dFlood and d99p of streamflow (i.e., the number of days of extreme streamflow exceeding the flooding and the 99th percentile thresholds) were moderately and positively correlated with leptospirosis cases. In comparison, the dnotFlood and d1p.d99p of streamflow (i.e., the number of days of average streamflow not exceeding the flooding threshold and occurring between the 1st and 99th percentile thresholds) were weakly and negatively correlated with leptospirosis cases. Water level indices exhibited very similar cross-correlation patterns as the streamflow indices.

The d90p of relative humidity (i.e., the number of days of extreme relative humidity exceeding the 90th percentile threshold) demonstrated positive correlations with leptospirosis cases at 7–13-week lag. Meanwhile, the d10p of relative humidity (i.e., the number of days of extreme relative humidity not exceeding the 10th percentile threshold) displayed positive correlations with leptospirosis cases at 21–52-week lag.

The d23 and d10p of temperature (i.e., the number of days of extreme temperature not exceeding the 23°C and 10th percentile thresholds) were positively correlated at shorter lags of up to 6 weeks. Meanwhile, the d34 and d90p of temperature (i.e., the number of days of extreme temperature exceeding the 34°C and 90th percentile thresholds) were positively correlated with leptospirosis cases at longer lags of 29–43 weeks.

Overall, extreme streamflow exhibited the highest correlation whereas average relative humidity exhibited the lowest correlation with leptospirosis cases (Table 3). The correlation of hydrometeorological variables was stronger under the extreme condition than the average condition. Most of the hydrometeorological extreme variables were positively correlated with leptospirosis cases compared with the average variables, which were negatively correlated with the disease.

Table 3 Summary of the value and direction of correlation between hydrometeorological indices and leptospirosis cases. + indicates a positive correlation, and – indicates negative correlation

Condition	Variable	Highest correlation	Short-term (1–13 weeks)	Long-term (27–52 weeks)
Average	Rainfall	0.29	+	–
	Streamflow/water level	0.30	–	Insignificant
	Relative humidity	0.26	Insignificant	–
	Temperature	0.30	–	–
Extreme	Rainfall	0.34	+	–
	Streamflow/water level	0.36	+	Insignificant
	Relative humidity	0.35	+	+
	Temperature	0.33	+	+

Table 4 The most and least important hydrometeorological indices to each preliminary model based on MDG

Model	Most important indices				Least important indices			
	Station	Variable	Index	MDG	Station	Variable	Index	MDG
A1	5721442	Water level	Mean	15.7	6122064	Rainfall	Mean	2.9
A2	6120014	Rainfall	d1mm.d50mm	11.1	5721442	Water level	dnotDanger	0.4
A3	6022001	Rainfall	d5p.d95p	9.9	5721442	Streamflow	d1p.d99p	0.4
A4	6120014	Rainfall	d1mm.d50mm	7.0	5721442	Water level	dDanger	0.1
E1	5920012	Rainfall	Maximum	19.9	6122064	Rainfall	Maximum	2.5
E2	5920012	Rainfall	d50mm	11.2	5721442	Streamflow	dFlood	0.1
E3	5920012	Rainfall	d95p	10.1	6019411	Streamflow	d99p	0.4
E4	5920012	Rainfall	d95p	5.9	5721442	Streamflow	dFlood	0.1
M1	5920012	Rainfall	Maximum	8.3	GK48617	Minimum temperature	Maximum	1.6
M2	5920012	Rainfall	d50mm	8.2	5721442	Streamflow	dFlood	0
M3	5920012	Rainfall	d95p	6.7	6019411	Streamflow	d99p	0.2
M4	5920012	Rainfall	d95p	3.4	5721442	Streamflow	dFlood	0
M23	5920012	Rainfall	d95p	12.9	6019411	Streamflow	d99p	0
M32	5920012	Rainfall	d50mm	11.6	5721442	Streamflow	dFlood	0.1

Model optimization based on MDG

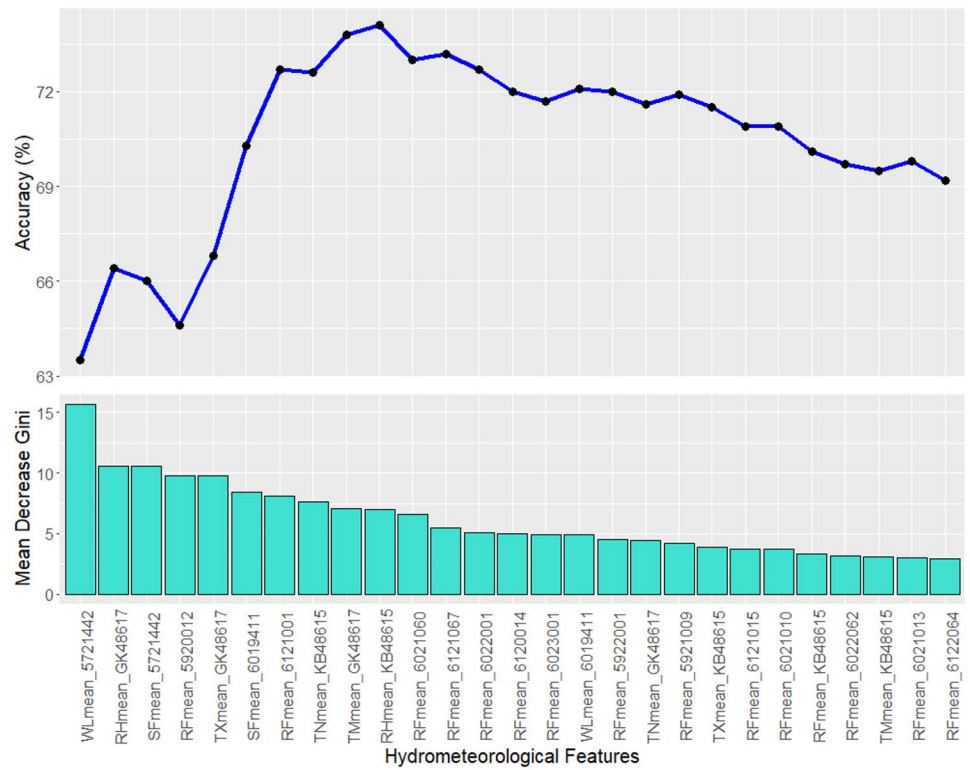
Table 4 shows the most and least important hydrometeorological indices to the preliminary models based on the mean decrease Gini (MDG) score. A5, E5, and M5 are excluded from Table 4 since MDG is biased towards the data with a large number of possible values, which could discriminate against the importance of indices with fewer possible values ranging from 0 to 7 days, i.e., relative and fixed days of hydrometeorological occurrence, that are features for these models. Overall, the rainfall indices appeared to have the highest importance in all models except in A1, whereas streamflow indices showed the least importance in most models including A3, E2, E3, E4, M2, M3, M4, M23, and M32 despite showing higher correlations with leptospirosis cases in “Cross-correlation analysis (CCA)”. In the mixed models, the extreme rainfall indices produced the highest MDG scores informing their high importance in classifying the leptospirosis occurrence.

The results of feature subset selection for A1 and M1 models are demonstrated in Fig. 2a and b respectively. The hydrometeorological features of the A1 model were ranked in descending order (from left to right in Fig. 2a) based on the MDG score obtained from the preliminary model. The most important feature was WLmean_5721442 since it exhibited the highest MDG (15.7) among the other features in the model. This feature was used to build the initial model. The initial model showed a training accuracy of 63.5%. Then, the accuracy increased from 63.5 to 66.4% when the initial model was incorporated with the second most important feature (RHmean_GK48617). However, the accuracy started to decrease when the weekly mean streamflow from station 5721442 and rainfall from station

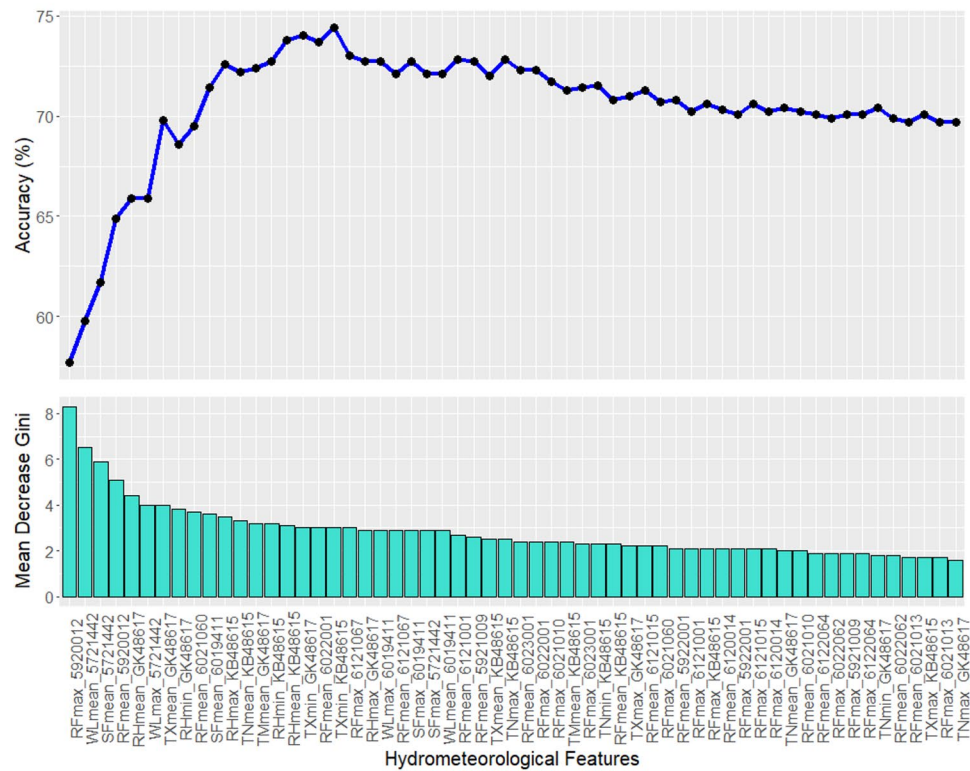
5920012 was added to the previous model. Then, the training accuracy increased from 64.6 to 74.1% with the consecutive inclusion of lower important mean indices, i.e., from TXmean_GK48617 to RHmean_KB48615 despite showing a slight reduction when TNmean_KB48615 was added to the previous subset. Afterwards, the accuracy deteriorated gradually from 74.1 to 69.2% as more indices with lower MDG values were added to the subset for model training. Similarly, Fig. 2b displays the feature subset selection for the M1 model. The training accuracy of the model increased with the addition of features from RFmax_5920012 (57.7%) to TXmin_KB48615 (74.4%) but started to decrease afterwards. These features were selected for the final model development.

The results of the retraining of models based on the MDG other than the A1 model are summarized in Fig. 3. All models managed to reduce their input data size by removing between three and 40 indices from the preliminary input dataset based on the ranking of MDG scores. The average number of features in the preliminary dataset was 44 but was reduced by 50% to 22 by the feature subset selection process. The only exception was A2 as it retained all the indices of its preliminary model. Nine to 12 models out of the 16 models showed an improvement in each metric after removing less important features based on MDG. On average, the training accuracy slightly increased from 67.0 to 68.9%, sensitivity increased from 37.0 to 39.3%, and specificity increased from 84.4 to 86.2%. The average specificity of models showed an acceptable range that lies between 80 and 90% while the average accuracy of the models was sufficient, which ranges from 60 to 70%. However, the average sensitivity of the models was below 50%, which could be due to the imbalance of output class in the

Fig. 2 a The training accuracy of A1 model with respect to increasing subset size of ordered training set based on MDG. **b** The training accuracy of M1 model with respect to increasing subset size of the ordered training set based on MDG



(a)



(b)

Fig. 3 Training performance metrics of models before and after selecting the reduced subsets and testing performance metrics of models with reduced subset



dataset. This resulted in the models performing poorly in detecting the high cases.

Predictive performance of models

The performance metrics were computed based on the optimal operating points selected from the ROC analysis. The ROC plot is presented in Figure S6 of Supporting Information. M1 (mixed model) showed the highest prediction accuracy of 82.6% indicating the ability of the model in correctly predicting both the high and low cases (Fig. 3). Meanwhile, A3 (average model) showed the lowest prediction accuracy of 61.5%. Generally, the sensitivity of the models was low. However, the E2 model displayed a sensitivity of 76.5%, which was the highest compared to the rest. This model was the most sensitive in correctly anticipating the high cases than the other models. The highest specificity of the A2, A3, and M1 (96.6%) models implied that the model was the most specific in correctly predicting the low cases than other models. When comparing between the average and extreme models, the most sensitive model (E2) was found among the extreme models while the most specific model (A2 and A3) was found among the average models. Overall, the average (61.5–76.1%) and extreme models (72.3–77.0%) showed similar prediction accuracy ranges while the mixed model (71.7–82.6%) showed an improvement.

Discussion

Cross-correlation analysis

The cross-correlations between hydrometeorological indices and leptospirosis cases appeared to be influenced by the monsoon seasons of the country. This finding supported previous studies that observed hydrometeorological factors to be critical to the seasonal development of leptospirosis (Joshi et al. 2017; Chadsuthi et al. 2012; Péres et al. 2019; Batchelor et al. 2012; Desvars et al. 2011).

The Malaysian climate showed distinctive control by two monsoons, i.e., the Northeast (November to March) and Southwest (May to September) monsoons. Usually, the former brings heavier rainfall, which often leads to flooding events in the east coast regions of the Peninsular Malaysia. In contrast, the Southwest monsoon is characterized by lower rainfall and higher temperatures. Positive correlations were observed between rainfall indices and leptospirosis up to the 15-week lag. During the wet season, leptospires can mobilize in the rain and floodwaters, subsequently infecting humans. The shorter time lag of 15 weeks could be related to their incubation period (Haake & Levett 2015). Additionally, it could account for the time taken for humans to be exposed to the bacteria. Similarly, a higher rainfall, a higher stream-flow and relative humidity, and a lower temperature were associated with higher leptospirosis cases at shorter lags.

Meanwhile, negative correlations were observed between rainfall indices and leptospirosis at a 35–52-week lag. Similar findings were reported in Chadsuthi et al. (2012). Correspondingly, a lower rainfall, a lower relative humidity, and a higher temperature were associated with higher leptospirosis cases at longer lags. This period could be related to the Southwest monsoon season, and the negative correlation could be due to a consequence of the higher number of cases found immediately after the Northeast monsoon. Furthermore, Hacker et al. (2020) have reported that although rainfall is strongly associated with leptospirosis, the disease tends to occur throughout the year. This could be due to other risk factors which could be indirectly related to hydrometeorological events. For instance, the increased mobility of humans during dry weather could lead to infection as they might interact with the environment contaminated by leptospires (Joshi et al. 2017).

Besides that, the positive correlation of extreme rainfall indices peaking at an earlier lag of 3 weeks, as compared to a latter lag for the corresponding average indices, was attributed to the immediate infection following flooding events caused by heavy rainfall. Floods could cause a more rapid infection by bringing leptospires closer to humans within a short period (Sehgal et al. 2002). The moderate positive correlation of extreme streamflow indices suggested that leptospirosis is better driven by extreme streamflow events compared to the average events. The latter are represented by the average indices, which demonstrate weaker correlations. Overbank inundation could pose a higher risk of infection among those residing in the vicinity of the river. Former studies observed that leptospirosis cases were more prevalent around rivers and when the river levels exceeded the danger limit; this was attributed to the river overflow that disperses leptospires towards the residential area (Hayati et al. 2018; López et al. 2019).

The correlation between simple indices of hydrometeorological variables and leptospirosis was similar for both average and extreme conditions. This is because the simple indices returned the mean and extremum (maximum and minimum) values for each week. To a certain extent, the mean contained information from the extreme values, as it redistributed the values among the number of days in the week. This is unlike the fixed and relative indices, which isolated the extreme values from the average values, and vice versa.

Lastly, the fixed thresholds used in this study may not have sufficiently represented the extreme limits of hydrometeorological events. The extreme thresholds tend to vary from one location to another depending on the regional climatic and geographical factors. Different levels of rainfall events would contribute to different levels of flooding at different locations.

Model optimization based on MDG

MDG tends to be biased towards the predictors with more possible values or categories compared to those with less possible values or categories in their data (Strobl et al. 2007). In this study, such biases were observed in the models (A5, E5, and M5) that incorporated both the indices with many and a few possible values. The higher the number of possible splits, the more often the index gets selected as the candidate predictor for the node split. Therefore, the MDG score of the frequently selected index tends to be larger as the score is summed up in each individual tree and averaged across the random forest.

Overall, the highest MDG scores of the rainfall indices in all models with the exception of A1 indicated that they contributed the most in decreasing the heterogeneity of nodes. This suggests the importance of rainfall in determining leptospirosis occurrence. This is in line with other studies that established a strong correlation between rainfall and leptospirosis (Hacker et al. 2020; Kupek et al. 2000; Cunha et al. 2022; Ghizzo Filho et al. 2018; Chadsuthi et al. 2012). On the other hand, the lowest MDG scores for the streamflow indices indicated that this variable contributed the least in decreasing the heterogeneity of nodes when growing the individual trees. This is despite streamflow having shown higher cross-correlations with leptospirosis. In cross-correlation analysis, each hydrometeorological index was analyzed individually assuming a one-to-one relationship between the features and leptospirosis. However, the MDG score of a hydrometeorological index was computed in consideration of other hydrometeorological indices during the model development. The rainfall feature would be frequently selected as the predictor for the nodes, which would reduce the participation of the streamflow feature in the model development. It may also be the case that there was less information from the streamflow time series which are more sparsely available as compared to the rainfall time series.

The less important hydrometeorological indices based on MDG can be characterized as noisy variables or variables with less information or possible values. The noisiness of features could be due to the random errors present in the data, which would lead to overfitting problems when the models attempt to conform closely to the existing data. The models tend to perform poorly when they see a different set of data that was not used during the model training. Apart from that, the less informative features would make the classification worse as there would not be sufficient possible values that would become the thresholds for splitting the outputs. However, subsetting the important features with decreasing MDG scores up to a certain number improved the model's accuracy since their combination recognized the signals better.

Predictive performance of models

The higher prediction accuracy of mixed models indicated that leptospirosis cases take place under the occurrence of both average and extreme hydrometeorological events. Usually, a higher number of cases (outbreak) occur during the extreme hydrometeorological events such as heavy rainfall and flood (Cann et al. 2013). Flooding events cause changes to human movements, which could ultimately reduce the distance and time of leptospirosis transmission. During normal (average) hydrometeorological events, humans are still infected with leptospirosis but probably lower in number (endemic) (Ghizzo Filho et al. 2018). The effect of average hydrometeorological events may not be as severe as extreme events. The socio-economic and rodent activities, which are influenced by the occurrence of hydrometeorological events, probably contribute to this infection.

Additionally, the similar ranges of prediction accuracy of the respective average and extreme models implied that these two conditions equally play their roles in the total leptospirosis occurrence. The normal hydrometeorological events might ensure the endemicity of leptospirosis, which records the lower number of cases throughout the year (Soo and Khan 2020). On the other hand, the extreme hydrometeorological events cause the disease to break out for a particular time (WHO 2001). Thus, the similar prediction accuracy ranges of average and extreme models reflected that the total number of cases has almost equal contributions by the respective average and extreme hydrometeorological indices.

The highest sensitivity of the extreme model showed that the extreme hydrometeorological indices incorporated in the model predicted the weeks with high cases better. Meanwhile, the highest specificity of the average model showed that the average hydrometeorological indices used to develop the model better predicted the weeks with low cases.

Conclusion

The ultimate aim of the work is to illustrate the importance of feature selection, i.e., the selection of the most relevant and non-redundant input, among average and extreme hydrometeorological indices in the construction of a classification model for leptospirosis. In response to the research objectives, which are to (1) identify the hydrometeorological indices highly cross-correlated with leptospirosis and important in classifying the disease occurrence and (2) to observe the prediction capacity change in response to the average, extreme indices, and their combination used as model features, the main conclusions are as follows:

- Rainfall was the most important, while streamflow was the least important variable based on mean decrease Gini despite showing higher cross-correlations with leptospirosis.
- Random forest models performed similarly with average and extreme hydrometeorological indices while their accuracy improved with both average and extreme hydrometeorological indices as features.
- The temporal lag between the hydrometeorological indices and leptospirosis followed the seasonality of the monsoon.

Future research can address the following research gaps. Firstly, a less biased approach such as the mean decrease accuracy (MDA) can be explored to measure the importance of hydrometeorological indices. A more reliable variable importance can help improve the model's accuracy. Agnostic machine learning methods such as SHapley Additive exPlanations (SHAP) (Lipovetsky and Conklin 2001) can be useful for this goal. Next, the analysis of extreme indices based on multiple thresholds that represent a varying severity of the hydrometeorological and their possible links to higher leptospirosis occurrences should be explored.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00484-022-02422-y>.

Acknowledgements We acknowledge the Department of Health Kelantan for providing access to the case data and the Department of Irrigation and Drainage Malaysia for providing the hydrological data. The authors would like to thank the Director General of Health Malaysia for the permission to publish this paper.

Author contribution Veianthan Jayaramu: conceptualization, methodology, software, formal analysis, investigation, data curation, writing—original draft, visualization, and project administration. Zed Zulkaffi: conceptualization, methodology, validation, writing—review and editing, visualization, supervision, project administration, and funding acquisition. Simon De Stercke: writing—review and editing, supervision, and project administration. Wouter Buytaert: writing—review and editing, project administration, and funding acquisition. Fariq Rahmat: software and data curation. Ribhan Zafira Abdul Rahman: writing—review and editing. Asnor Juraiza Ishak: writing—review and editing. Wardah Tahir: writing—review and editing. Jamalludin Ab Rahman: writing—review and editing. Nik Mohd Hafiz Mohd Fuzi: resources and writing—review and editing.

Funding This work was supported by grants from the Ministry of Higher Education Malaysia (NEWTON/1/2018/WAB05/UPM/1) and from the UK Natural Research Environment Council (NE/S003053/1) under the Understanding of the Impacts of Hydrometeorological Hazards in South East Asia program.

Declarations

Ethical approval Ethical approval for this study was obtained from the Medical Research and Ethics Committee, Ministry of Health Malaysia (NMRR-19-4115-47702).

Conflict of interest The authors declare no competing interests.

References

- Adler B, de la Peña Moctezuma A (2010) *Leptospira* and leptospirosis. *Vet Microbiol* 140(3–4):287–296. <https://doi.org/10.1016/j.vetmic.2009.03.012>
- Ahangarcani M, Farnaghi M, Shirzadi MR, Pilesjö P, Mansourian A (2019) Predictive risk mapping of human leptospirosis using support vector machine classification and multilayer perceptron neural network. *Geospatial Health* 14(1). <https://doi.org/10.4081/gh.2019.711>
- Ansdell VE (2017) Chapter 23 - Leptospirosis. Elsevier Inc., In *The Travel and Tropical Medicine Manual (Fifth Edition)*, pp 336–344. <https://doi.org/10.1016/B978-0-323-37506-1.00023-4>
- Barradas-Bautista D (2020) Random forest and deep learning performance on the Malaria DREAM sub challenge one random forest and deep learning performance. *Res Comput Sci* 149(5):163–170
- Benacer D, Thong KL, Min NC, Verasahib KB, Galloway RL, Hartskeerl RA, Souris M, Zain SNM (2016) Epidemiology of human leptospirosis in Malaysia, 2004–2012. *Acta Tropica* 157:162–168. <https://pubmed.ncbi.nlm.nih.gov/26844370/>
- Barcellos C, Sabroza PC (2001) The place behind the case: leptospirosis risks and associated environmental conditions in a flood-related outbreak in Rio de Janeiro. *Cad Saúde Pública / Ministério Da Saúde, Fundação Oswaldo Cruz, Escola Nacional De Saúde Pública* 17(Suppl):59–67. <https://doi.org/10.1590/s0102-311x2001000700014>
- Batchelor TWK, Stephenson TS, Brown PD, Amarakoon D, Taylor MA (2012) Influence of climate variability on human leptospirosis cases in Jamaica. *Climate Res* 55(1):79–90. <https://doi.org/10.3354/cr01120>
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
- Campbell AM, Racault MF, Goult S, Laurenson A (2020) Cholera risk: a machine learning approach applied to essential climate variables. *Int J Environ Res Public Health* 17(24):1–24. <https://doi.org/10.3390/ijerph17249378>
- Cann KF, Thomas DR, Salmon RL, Wyn-Jones AP, Kay D (2013) Extreme water-related weather events and waterborne disease. *Epidemiol Infect* 141(4):671–686. <https://doi.org/10.1017/S0950268812001653>
- Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K (2018) Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila Philippines. *BMC Infect Dis* 18(1):1–15. <https://doi.org/10.1186/s12879-018-3066-0>
- Chadsuthi S, Modchang C, Lenbury Y, Iamsirithaworn S, Triampo W (2012) Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses. *Asian Pac J Trop Med* 5(7):539–546. [https://doi.org/10.1016/S1995-7645\(12\)60095-9](https://doi.org/10.1016/S1995-7645(12)60095-9)
- Cunha M, Costa F, Ribeiro GS, Carvalho MS, Reis RB, Nery Jr N, Pischel L, Gouveia EL, Santos AC, Queiroz A, Wunder Jr EA, Reis MG, Diggle PJ, Ko AI (2022) Rainfall and other meteorological factors as drivers of urban transmission of leptospirosis. *PLoS Negl Trop Dis* 16(4):e0007507. <https://doi.org/10.1101/658872>
- Department of Irrigation and Drainage Malaysia [DID] (2017) Flood management - programme and activities. <https://www.water.gov.my/index.php/pages/view/419?mid=244>. Accessed 27 Nov 2020
- Desvars A, Jégo S, Chiroleu F, Bourhy P, Cardinale E, Michault A (2011) Seasonality of human leptospirosis in Reunion Island (Indian Ocean) and its association with meteorological data. *PLoS ONE* 6(5):e20377. <https://doi.org/10.1371/journal.pone.0020377>
- Dhewantara PW, Lau CL, Allan KJ, Hu W, Zhang W, Mamun AA, Soares Magalhães RJ (2019) Spatial epidemiological approaches to inform leptospirosis surveillance and control: A systematic review and critical appraisal of methods. *Zoonoses and Public Health* 66(2):185–206. <https://doi.org/10.1111/zph.12549>
- Ding G, Li X, Li X, Zhang B, Jiang B, Li D, Xing W, Liu Q, Liu X, Hou H (2019) A time-trend ecological study for identifying flood-sensitive infectious diseases in Guangxi, China from 2005 to 2012. *Environ Res* 176(July):108577. <https://doi.org/10.1016/j.envres.2019.108577>
- Ehelepola NDB, Ariyaratne K, Dissanayake WP (2019) The correlation between local weather and leptospirosis incidence in Kandy district, Sri Lanka from 2006 to 2015. *Global Health Action* 12(1):1553283. <https://doi.org/10.1080/16549716.2018.1553283>
- Ghizzo Filho J, Nazário NO, Freitas PF, Pinto GDA, Schlindwein AD (2018) Temporal analysis of the relationship between leptospirosis, rainfall levels and seasonality, Santa Catarina, Brazil, 2005–2015. *Rev Inst Med Trop Sao Paulo* 3154(01):18–17. <https://doi.org/10.1590/S1678-9946201860039>
- Glaros AG, Kline RB (1988) Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *J Clin Psychol* 44(6):1013–1023. [https://doi.org/10.1002/1097-4679\(198811\)44:6%3C1013::AID-JCLP2270440627%3E3.0.CO;2-Z](https://doi.org/10.1002/1097-4679(198811)44:6%3C1013::AID-JCLP2270440627%3E3.0.CO;2-Z)
- Gómez AA, López MS, Müller GV, López LR, Sione W, Giovannini L (2022) Modeling of leptospirosis outbreaks in relation to hydroclimatic variables in the northeast of Argentina. *Heliyon* 8(6):e09758. <https://doi.org/10.1016/j.heliyon.2022.e09758>
- Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, Luo G, Li Z, He J, Zhang Y, Ma W (2017) Developing a dengue forecast model using machine learning: a case study in China. *PLoS Negl Trop Dis* 11(10):e0005973. <https://doi.org/10.1371/journal.pntd.0005973>
- Haake DA, Levett PN (2015) Leptospirosis in humans. *Curr Top Microbiol Immunol* 387:65–97. https://doi.org/10.1007/978-3-662-45059-8_5
- Hacker KP, Sacramento GA, Cruz JS, De Oliveira D, Nery N, Lindow JC, Carvalho M, Hagan J, Diggle PJ, Begon M, Reis MG, Wunder EA, Ko AI, Costa F (2020) Influence of rainfall on leptospira infection and disease in a tropical urban setting Brazil. *Emerg Infect Dis* 26(2):311–314. <https://doi.org/10.3201/eid2602.190102>
- Hayati KS, Sharifah Norkhadajah SI, Salmiah MS, Edre MA, Khin TD (2018) Hot-spot and cluster analysis on legal and illegal dumping sites as the contributors of leptospirosis in a flood hazard area in Pahang, Malaysia. *Asian J Agric Biol* 5(2):56–59
- Hu H, Wang H, Wang F, Langley D, Avram A, Liu M (2018) Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. *Sci Rep* 8(1):1–8. <https://doi.org/10.1038/s41598-018-23075-1>
- Ismail WR, Haghroosta T (2018) Extreme weather and floods in Kelantan state, Malaysia in December 2014. *Res Mar Sci* 3(1):231–244
- Jamaludin N, Mohammed NI, Khamidi MF, Wahab SNA (2015) Thermal comfort of residential building in Malaysia at different microclimates. *Proc Soc Behav Sci* 170:613–623. <https://doi.org/10.1016/j.sbspro.2015.01.063>
- Joshi YP, Kim EH, Cheong HK (2017) The influence of climatic factors on the development of hemorrhagic fever with renal syndrome and leptospirosis during the peak season in Korea: an ecologic study. In *BMC Infect Dis* 17(1). <https://doi.org/10.1186/s12879-017-2506-6>
- Khan S, Ullah R, Khan A, Sohail A, Wahab N, Bilal M, Ahmed M (2017) Random forest-based evaluation of Raman spectroscopy for dengue fever analysis. *Appl Spectrosc* 71(9):2111–2117. <https://doi.org/10.1177/0003702817695571>
- Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Soft* 28(5):1–26. <https://doi.org/10.18637/jss.v028.i05>
- Kumar V, Minz S (2014) Feature selection: a literature review. *Smart Comput Rev* 4(3):211–229. <https://doi.org/10.6029/smartcr.2014.03.007>

- Kupek E, de Sousa Santos Faversoni MC, de Souza Philippi JM (2000) The relationship between rainfall and human leptospirosis in Florianópolis, Brazil, 1991–1996. *Braz J Infect Dis: Off Publ Braz Soc Infect Dis* 4(3):131–134
- Lau CL, Clements ACA, Skelly C, Dobson AJ, Smythe LD, Weinstein P (2012) Leptospirosis in American Samoa - estimating and mapping risk using environmental data. *PLoS Negl Trop Dis* 6(5):e1669. <https://doi.org/10.1371/journal.pntd.0001669>
- Lau CL, Smythe LD, Craig SB, Weinstein P (2010) Climate change, flooding, urbanisation and leptospirosis: fuelling the fire? *Trans R Soc Trop Med Hyg* 104(10):631–638. <https://doi.org/10.1016/j.trstmh.2010.07.002>
- Lipovetsky S, Conklin M (2001) Analysis of regression in game theory approach. *Appl Stoch Model Bus Ind* 17(4):319–330. <https://doi.org/10.1002/asmb.446>
- López MS, Müller GV, Lovino MA, Gómez AA, Sione WF, Aragonés Pomares L (2019) Spatio-temporal analysis of leptospirosis incidence and its relationship with hydroclimatic indicators in northeastern Argentina. *Sci Total Environ* 694. <https://doi.org/10.1016/j.scitotenv.2019.133651>
- Mayfield HJ, Lowry JH, Watson CH, Kama M, Nilles EJ, Lau CL (2018) Use of geographically weighted logistic regression to quantify spatial variation in the environmental and sociodemographic drivers of leptospirosis in Fiji: a modelling study. *Lancet Planet Health* 2(5):e223–e232. [https://doi.org/10.1016/S2542-5196\(18\)30066-4](https://doi.org/10.1016/S2542-5196(18)30066-4)
- Mohammadinia A, Alimohammadi A, Saeidian B (2017) Efficiency of geographically weighted regression in modeling human leptospirosis based on environmental factors in Gilan province Iran. *Geosci* 7(4):136. <https://doi.org/10.3390/geosciences7040136>
- Mohd Radi MF, Hashim JH, Jaafar MH, Hod R, Ahmad N, Nawi AM, Baloch GM, Ismail R, Ayub NIF (2018) Leptospirosis outbreak after the 2014 major flooding event in Kelantan, Malaysia: a spatial-temporal analysis. *Am J Trop Med Hyg* 98(5):1281–1295. <https://doi.org/10.4269/ajtmh.16-0922>
- Péres WE, Russo A, Nunes B (2019) The association between hydro-meteorological events and leptospirosis hospitalizations in Santa Catarina Brazil. *Water* 11(5):1052. <https://doi.org/10.3390/w11051052>
- Peterson TC, Folland CC, Gruza G, Hogg W, Mokssit A, Plummer N (2001) Report on the activities of the working group on climate change detection and related rapporteurs 1998–2001. Rep. WCDMP-47, WMO-TD 1071, Geneva, Switzerland, March, 143. <http://etccdi.pacificclimate.org/docs/wgccd.2001.pdf>. Accessed 10 Oct 2022
- Picardeau M (2013) Diagnosis and epidemiology of leptospirosis. *Med Et Mal Infect* 43(1):1–9. <https://doi.org/10.1016/j.medmal.2012.11.005>
- Rahayu S, Adi MS, Saraswati LD (2018) Mapping of leptospirosis environmental risk factors and determining the level of leptospirosis vulnerable zone in Demak District using remote sensing image. In: *E3S Web of Conferences 2018*. EDP Sciences, vol. 31, p 06003. <https://doi.org/10.1051/e3sconf/20183106003>
- Rahmat F, Ishak AJ, Zulkafli Z, Yahaya H, Masrani A (2019) Prediction model of leptospirosis occurrence for Seremban (Malaysia) using meteorological data. *Int J Integr Eng* 11(4):61–69. <https://doi.org/10.30880/ijie.2019.11.04.007>
- Rahmat F, Zulkafli Z, Ishak AJ, Mohd Noor SB, Yahaya H, Masrani A (2020) Exploratory data analysis and artificial neural network for prediction of leptospirosis occurrence in Seremban, Malaysia Based on Meteorological Data. *Front Earth Sci* 8:377. <https://doi.org/10.3389/feart.2020.00377>
- Sánchez-Montes S, Espinosa-Martínez DV, Ríos-Muñoz CA, Berzunza-Cruz M, Becker I (2015) Leptospirosis in Mexico: epidemiology and potential distribution of human cases. *PLoS ONE* 10(7):e0133720. <https://doi.org/10.1371/journal.pone.0133720>
- Santos MS, Soares JP, Abreu PH, Araujo H, Santos J (2018) Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. *IEEE Comput Intell Mag* 13(4):59–76. <https://doi.org/10.1109/MCI.2018.2866730>
- Schneider MC, Nájera P, Aldighieri S, Bacallao J, Soto A, Marquiño W, Altamirano L, Saenz C, Marin J, Jimenez E, Moynihan M, Espinal M (2012) Leptospirosis outbreaks in Nicaragua: identifying critical areas and exploring drivers for evidence-based planning. *Int J Environ Res Public Health* 9(11):3883–3910. <https://doi.org/10.3390/ijerph9113883>
- Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: appropriate use and interpretation. *Anesth Analg* 126(5):1763–1768. <https://doi.org/10.1213/ANE.0000000000002864>
- Sehgal SC, Sugunan AP, Vijayachari P (2002) Outbreak of leptospirosis after the cyclone in Orissa. *Natl Med J India* 15(1):22–23
- Soo ZMP, Khan NA, Siddiqui R (2020) Leptospirosis: increasing importance in developing countries. *Acta Tropica* 201:105183. <https://doi.org/10.1016/j.actatropica.2019.105183>
- Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinforma* 8(1):1–21. <https://doi.org/10.1186/1471-2105-8-25>
- Sumi A, Telan EFO, Chagan-Yasutan H, Piolo MB, Hattori T, Kobayashi N (2017) Effect of temperature, relative humidity and rainfall on dengue fever and leptospirosis infections in Manila, the Philippines. *Epidemiol Infect* 145(1):78–86. <https://doi.org/10.1017/S095026881600203X>
- Suwanpakdee S, Kaewkungwal J, White LJ, Asensio N, Ratanakorn P, Singhasivanon P, Day NPJ, Pan-Ngum W (2015) Spatio-temporal patterns of leptospirosis in Thailand: is flooding a risk factor? *Epidemiol Infect* 143(10):2106–2115. <https://doi.org/10.1017/S0950268815000205>
- Tassinari WS, Pellegrini DC, Sá CB, Reis RB, Ko AI, Carvalho MS (2008) Detection and modelling of case clusters for urban leptospirosis. *Tropical Med Int Health* 13(4):503–512. <https://doi.org/10.1111/j.1365-3156.2008.02028.x>
- Togami E, Kama M, Goarant C, Craig SB, Lau C, Ritter JM, Imrie A, Ko AI, Nilles EJ (2018) A large leptospirosis outbreak following successive severe floods in Fiji, 2012. *American Journal of Tropical Medicine and Hygiene*, 99(4), 849–851. <https://doi.org/10.4269/ajtmh.18-0335>
- Ucar MK, Nour M, Sindi H, Polat K (2020) The effect of training and testing process on machine learning in biomedical datasets. *Math Prob Eng* 2020:17. <https://doi.org/10.1155/2020/2836236>
- Uusitalo R, Siljander M, Dub T, Sane J, Sormunen JJ, Pellikka P, Vapalahti O (2020) Modelling habitat suitability for occurrence of human tick-borne encephalitis (TBE) cases in Finland. *Ticks and Tick-borne Diseases* 11(5):101457. <https://doi.org/10.1016/j.ttbdis.2020.101457>
- Van Stralen KJ, Stel VS, Reitsma JB, Dekker FW, Zoccali C, Jager KJ (2009) Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. *Kidney Int* 75(12):1257–1263. <https://doi.org/10.1038/ki.2009.92>
- Vega-Corredor M, Opadeyi J (2014) Hydrology and public health: linking human leptospirosis and local hydrological dynamics in Trinidad, West Indies. *Earth Perspectives* 1:1–4. <https://doi.org/10.1186/2194-6434-1-3>
- Weinberger D, Baroux N, Grangeon JP, Ko AI, Goarant C (2014) El Niño southern oscillation and leptospirosis outbreaks in New Caledonia. *PLoS Negl Trop Dis* 8(4):e2798. <https://doi.org/10.1371/journal.pntd.0002798>
- World Health Organization (2001) WHO recommended strategies for the prevention and control of communicable diseases (No. WHO/CDS/CPE/SMT/2001.13). World Health Organization. <https://>

- apps.who.int/iris/bitstream/handle/10665/67088/WHO_CDS_CPE_SMT_2001.13.pdf. Accessed 22 May 2021
- World Health Organization (2003) Human leptospirosis: guidance for diagnosis, surveillance and control (No. WHO/CDS/CSR/EPH 2002.23). World Health Organization. <https://www.who.int/publications/i/item/human-leptospirosis-guidance-for-diagnosis-surveillance-and-control>. Accessed 17 Aug 2021
- World Health Organization (2011) Report of the Second Meeting of the Leptospirosis Burden Epidemiology Reference Group. World Health Organization. http://apps.who.int/iris/bitstream/handle/10665/44588/9789241501521_eng.pdf?sequence=1. Accessed 22 Aug 2021
- Zakharova OI, Korennoy FI, Iashin IV, Toropova NN, Gogin AE, Kolbasov DV, Surkova GV, Malkhazova SM, Blokhin AA (2021) Ecological and Socio-economic determinants of livestock animal leptospirosis in the Russian arctic. *Front Vet Sci* 8:658675. <https://doi.org/10.3389/fvets.2021.658675>
- Zhang X, Alexander L, Hegerl GC, Jones P, Tank AK, Peterson TC, Trewin B, Zwiers FW (2011) Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdiscip Rev: Clim Chang* 2(6):851–870. <https://doi.org/10.1002/wcc.147>
- Zhang Z, Yang Z, Ren W, Wen G (2019) Random forest-based real-time defect detection of Al alloy in robotic arc welding using optical spectrum. *J Manuf Process* 42:51–59. <https://doi.org/10.1016/j.jmapro.2019.04.023>
- Zhao J, Liao J, Huang X, Zhao J, Wang Y, Ren J, Wang X, Ding F (2016) Mapping risk of leptospirosis in China using environmental and socioeconomic data. *BMC Infect Dis* 16(1):1–10. <https://doi.org/10.1186/s12879-016-1653-5>
- Zhao N, Charland K, Carabali M, Nsoesie EO, Maheu-Giroux M, Rees E, Yuan M, Garcia Balaguera C, Jaramillo Ramirez G, Zinszer K (2020) Machine learning and dengue forecasting: Comparing random forest and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLoS Negl Trop Dis* 14(9):e0008056. <https://doi.org/10.1371/journal.pntd.0008056>
- Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.