Spatio-Temporal Prediction of Leptospirosis in Kuantan Using Hydrometeorological Variables and Random Forest Machine Learning

Veianthan Jayaramu
Department of Civil Engineering, Universiti Putra Malaysia

Zed Zulkafli
Department of Civil Engineering, Universiti Putra Malaysia

Fariq Rahmat
Department of Electrical and Electronic Engineering, Universiti Putra Malaysia

https://doi.org/10.5109/7323369

出版情報：Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES). 10, pp.916-922, 2024-10-17. International Exchange and Innovation Conference on Engineering & Sciences
バージョン：

Spatio-Temporal Prediction of Leptospirosis in Kuantan Using Hydrometeorological Variables and Random Forest Machine Learning

Veianthan Jayaramu[1,3], Zed Zulkafli[1], Fariq Rahmat[2]
[1]Department of Civil Engineering, Universiti Putra Malaysia, Serdang, Malaysia, [2]Department of Electrical and Electronic Engineering, Universiti Putra Malaysia, Serdang, Malaysia, [3]Ocned Water Technology Sdn. Bhd.
Corresponding author email: zeddiyana@upm.edu.my

Abstract: *Leptospirosis, a zoonotic disease prevalent in tropical regions with consistent rainfall, has been extensively studied using hydrometeorological data. This study focuses on developing a spatio-temporal model to predict leptospirosis in the Kuantan district, Pahang, known for its heavy rainfall and high disease incidence. Utilizing the random forest machine learning algorithm, we integrated hydrometeorological variables such as rainfall, streamflow, water level, relative humidity, and temperature across four model scenarios, lagging them from zero to 12 weeks at four-weeks intervals. Our models achieved an average testing accuracy of 73.4%, with sensitivity and specificity of 83.8% and 62.9%, respectively. Notably, we observed a minimal variation among the model scenarios, contrasting with previous studies where lag time improved the results. These findings underscore the potential of our models as a predictive tool for leptospirosis, enhancing spatial and temporal understanding in the Kuantan district. This improved insight can inform targeted disease prevention strategies, ultimately aiding in better management of leptospirosis outbreaks.*

Keywords: Leptospirosis, Hydrometeorological Variables, Spatio-Temporal Modelling, Disease Prediction, Random Forest Machine Learning

1. INTRODUCTION

Leptospirosis is a zoonotic disease caused by *Leptospira* sp. bacteria, primarily carried by rodents such as rats [1]. It is globally associated with high morbidity and mortality rates, particularly in developing countries with inadequate sanitation and health infrastructure [2]. The prevalence of leptospirosis is notably higher in tropical and subtropical regions, attributable to various risk factors inherent to these [3]. One significant factor is the frequent rainfall events, which play a critical role in the transmission dynamics of the disease. Rainfall facilitates the movement of Leptospira from rodents into the environment, increasing the risk of human exposure.

Tropical countries, characterized by consistent rainfall throughout the year, often report higher incidences of leptospirosis. Consequently, rainfall data has been extensively utilized in mapping the temporal occurrence of the disease. Numerous modelling studies have been conducted to understand the patterns and drivers of leptospirosis outbreaks [4]. Temporal models have been developed to analyse disease trends over time, while spatial models have been employed to examine geographic distribution. To achieve a comprehensive understanding of leptospirosis at a local level, spatio-temporal models have also been created, integrating both temporal and spatial dimensions.

Kuantan, a district in the Pahang state on the East Coast of Peninsular Malaysia, presents a unique case study for leptospirosis due to its distinctive climatic and environmental conditions [5]. Despite Malaysia's overall tropical climate, the East Coast districts, including Kuantan, experience significantly higher rainfall, especially during the Northeast Monsoon season. This increased precipitation creates a conducive environment for the spread of waterborne diseases such as leptospirosis.

Over the past decade, from 2011 to 2020, Kuantan has recorded a notably higher number of leptospirosis cases compared to other districts of Pahang. The frequent and intense rainfall in Kuantan leads to water accumulation and flooding, creating ideal conditions for the bacteria to thrive and spread. Consequently, the district has become a focal point for public health concerns regarding leptospirosis.

Recurring leptospirosis outbreaks in Kuantan significantly strain public health resources and affect community well-being. The economic burden is substantial, encompassing both direct medical costs and indirect costs such as loss of productivity due to illness and prolonged recovery periods. The agricultural and tourism sectors, essential to the local economy, also suffer during these outbreaks, causing broader socio-economic repercussions. Understanding the spatio-temporal dynamics of leptospirosis in relation to hydrometeorological variables is therefore crucial for developing effective prevention and control strategies, ultimately reducing the disease burden and associated economic impacts.

Given the rising incidence rates and the environmental factors at play, this research aims to develop an effective prediction model for leptospirosis in Kuantan using hydrometeorological factors as independent variables and employing the random forest machine learning to analyse the data from a spatio-temporal perspective, and make accurate predictions of leptospirosis occurrence over the study area. The main objectives to achieve this aim are:

1. To assess and analyse the hydrometeorological data for reliability;
2. To transform the input data consisting of both hydrometeorological and leptospirosis data into a spatio-temporal format;
3. To train the model with 60% of the data, and test the trained model with the remaining 40% of unseen data.

The novelty of this research lies in focusing on Kuantan's unique hydrometeorological conditions, developing a spatio-temporal prediction model for leptospirosis using the random forest technique, and highlighting the

importance of aiding healthcare authorities in implementing timely and targeted precautionary measures. By accurately predicting potential outbreaks, this proactive approach could significantly enhance the health and well-being of the local population, underscoring the importance of this research in combating leptospirosis in Kuantan.

2. MATERIALS AND METHODS

This section details the research methodology employed in this study, elaborating on the various methods and techniques used to conduct the investigation.

2.1 Study area

Kuantan, a district in Pahang, serves as the capital city of the state. Situated on the East Coast of Peninsular Malaysia, Kuantan is highly susceptible to severe flooding events, primarily due to heavy and persistent rainfall during the Northeast Monsoon season. Covering an area of approximately 2,960 square kilometers, the district features a tropical rainforest climate, marked by significant rainfall throughout the year. The hydrological landscape of Kuantan is influenced by several major rivers, notably the Kuantan River and parts of the Pahang River. These rivers frequently overflow during periods of intense rainfall, exacerbating the region's flooding problems. This study focuses on this particular area, as depicted in Fig. 1, to develop a spatio-temporal model to predict the occurrence of leptospirosis.

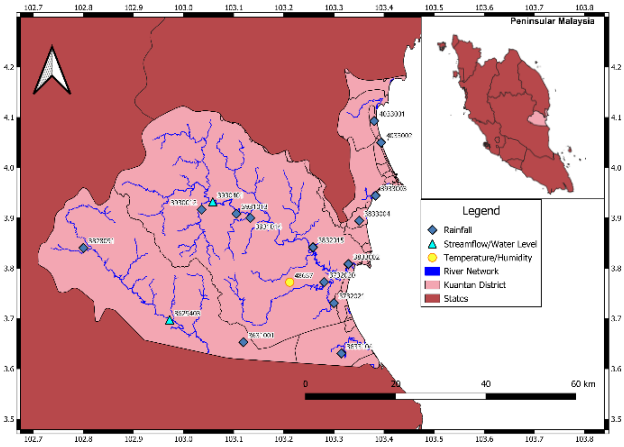

Fig. 1. Study area and locations of hydrometeorological stations.

2.2 Data collection

From 2011 to 2020, the Kuantan district had a higher incidence rate of 146.6 per 100,000 population, recording 625 cases, the highest among all districts, according to the e-notification data retrieved from the Pahang State Health Department. This period saw a significant rise in leptospirosis cases, highlighting Kuantan as the most affected district in the region. The data from the e-notification system provides a comprehensive overview of the public health landscape, emphasizing the severity of the leptospirosis outbreak in Kuantan compared to other districts. The high incidence rate of 146.6 per 100,000 population and the total of 625 recorded cases underscore the critical need for targeted public health interventions and robust disease management strategies in the Kuantan district.

The spatial distribution of hydrometeorological stations used for this study is shown in Fig. 1, while the list of data utilized is presented in Table 1. The study area's hydrometeorological data, encompassing rainfall, streamflow, water level, relative humidity, and temperature, were sourced from two primary agencies: the Department of Irrigation and Drainage Malaysia (DID) and the Malaysian Meteorological Department (MetMalaysia). The DID provides extensive data on water-related parameters, such as rainfall, streamflow, and water levels, which are vital for understanding the dynamics of river systems and potential flooding events in the region. Meanwhile, the MetMalaysia supplies detailed meteorological data, including rainfall, relative humidity, and temperature readings. This comprehensive meteorological data is essential for assessing climatic conditions and their impact on the hydrological cycle.

Table 1. List of hydrometeorological data used.

Data Type	Station ID	Source
Rainfall	RF3631001	DID
	RF3633104	
	RF3732020	
	RF3732021	
	RF3828091	
	RF3832015	
	RF3833002	
	RF3833004	
	RF3930012	
	RF3931013	
	RF3931014	
	RF3933003	
	RF4033001	
	RF4033002	
	RF48657	MetMalaysia
Streamflow	SF3930401	DID
	SF3629403	
Water Level	WL3930401	
	WL3629403	
Relative Humidity	RH48657	MetMalaysia
Temperature	TX48657	
	TM48657	
	TN48657	

2.3 Data processing

The precise case locations where the disease was contracted were identified by analysing comments within the e-notification dataset provided by the Pahang State Health Department. These locations were then mapped spatially using QGIS software to visualize the distribution of cases across the district. To facilitate a detailed spatio-temporal analysis, a grid measuring 5 km by 5 km was overlaid on the entire study area, as shown in Fig. 2. This grid structure allowed for the production of time series data on a weekly basis for each grid cell.

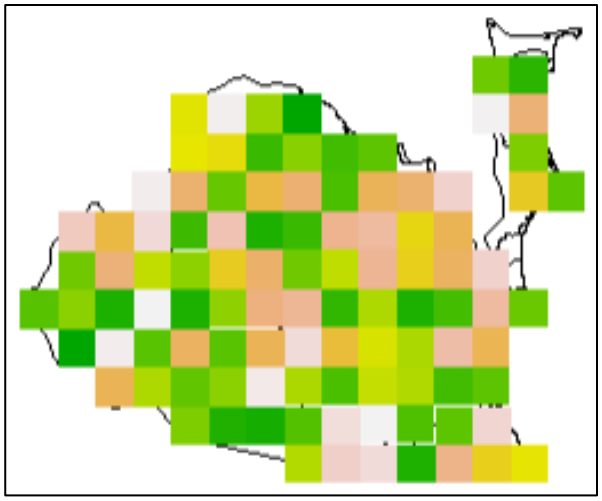

Fig. 2. 5 km by 5 km grid generated over the study area.

By segmenting the study area into these smaller units, the changes in disease incidence over time and across different locations could be analysed, providing a more granular understanding of the spatio-temporal dynamics of leptospirosis in Kuantan. By mapping the incidence data within these grid cells, patterns of spatial distribution could be discerned, providing insights into environmental and geographical factors contributing to the spread of leptospirosis. The use of QGIS software ensured accurate geospatial representation and analysis, supporting the development of targeted public health interventions and resource allocation strategies aimed at mitigating the impact of the disease in Kuantan.

The case data were meticulously organized into a time series for each grid cell and aggregated on a weekly basis. The daily hydrometeorological data, including rainfall, streamflow, water level, relative humidity, and temperature, were averaged weekly to create a consistent temporal dataset. This weekly averaging process ensured that short-term fluctuations did not obscure longer-term trends.

The rainfall data were subjected to spatial interpolation using the Kriging method [6], a geostatistical technique that provides accurate and reliable spatial estimates. This interpolation was performed for each weekly time step, and the results were then cropped into 5 km by 5 km grids. This process resulted in a comprehensive time series of rainfall data for each grid cell, capturing spatial variations in precipitation across the study area.

The relative humidity and temperature data, on the other hand, were used without spatial interpolation. This decision was based on the availability of a single relevant meteorological station for these variables within the study area, making spatial interpolation unnecessary.

The study area, dominated by the Pahang River and Kuantan River catchments, required specific attention to streamflow and water level data. Grids located within these catchments were assigned corresponding to the streamflow and water level data, ensuring that the hydrological dynamics of these important river systems were accurately represented in the model.

2.4 Data analysis

Several data analyses were conducted to check the quantity and quality of the hydrometeorological data used in this study. These analyses included checking data availability, consistency, reliability and redundancy.

First, data availability was assessed by calculating the percentage of missing data. This step involved a thorough examination of each dataset to identify any gaps or missing values. By quantifying the extent of missing data, we could determine the completeness of the dataset and make necessary adjustments to ensure robust analysis.

Next, we evaluated data consistency by plotting streamflow against water level data. This involved creating scatter plots to visually inspect the relationship between these two variables. Consistent data would show a clear, logical correlation between streamflow and water level, indicating that the measurements were accurate and reliable.

To assess data reliability, we produced single mass curves [7] for each of the rainfall stations. The single mass curve is a cumulative plot of rainfall data over time, which helps identify any inconsistencies or anomalies in the data. A smooth and continuous curve indicates reliable data, while deviations from this pattern may suggest errors or irregularities in the measurements.

Finally, we addressed data redundancy by conducting a multicollinearity test [8] between minimum, maximum, and mean temperature data received from the MetMalaysia. Multicollinearity occurs when independent variables in a dataset are highly correlated, which can distort the results of statistical analyses. By testing for multicollinearity, we ensured that our daily minimum temperature data were not redundant to the model.

2.5 Input data preparation

The study area is composed of 85 grid cells, each measuring 5 km by 5 km. Of these, the Kuantan River Basin encompassed 53 grid cells, while the Pahang River Basin contained 32 grid cells. Weekly processed data on leptospirosis cases and hydrometeorological variables were extracted for each grid cell. Each grid cell included five weekly hydrometeorological time series as independent variables: rainfall, streamflow, water level, relative humidity, and temperature. Additionally, one weekly time series for leptospirosis cases served as the dependent variable. For constructing the training dataset, 60% of the data from each grid cell was compiled across all cells, ensuring a comprehensive representation of the study area. To facilitate the analysis, weeks with one or more reported cases of leptospirosis were classified as "Yes," indicating the presence of the disease, while weeks with no reported cases were classified as "No." This binary classification allowed for the application of machine learning techniques to predict the occurrence of leptospirosis based on the hydrometeorological conditions. The data structure of the input data is shown in Fig. 3.

Date	Rainfall	Level	Flow	Humidity	Temperature	Cases
16/6/2013	0.764865518	16.35857143	0.672857143	80.27142857	25.2	No
27/9/2020	5.088587672	15.87714286	0.025714286	80.28571429	23.61428571	No
5/1/2020	1.086246464	16.01834101	1.854285714	80.45714286	23.55714286	No
18/9/2016	7.747346856	16.43	0.948571429	86.84285714	23.6	No
28/2/2016	3.589645568	16.43857143	0.975714286	81.11428571	23.37142857	No
27/7/2014	4.808505832	16.49714286	1.191428571	79.88571429	24.21428571	No
13/11/2016	7.112615777	17.28428571	52.09285714	89.61428571	23.81428571	No
15/3/2020	0.661417372	15.75571429	0.007142857	79.74285714	24.32857143	No
12/8/2012	2.734823495	17.2697786	63.0304797	82.6	23.64285714	No
17/4/2016	2.009255907	16.57142857	1.488571429	82.25714286	25.64285714	No

Fig. 3. Structure of input data prepared for machine learning.

2.6 Data partitioning

The input data for the model was divided into two sets: 60% for the training set and 40% for the testing set. Allocating a higher percentage to the training set is a common practice as was also performed by [9] to ensure that the model has ample data to learn from, thereby improving its accuracy and performance [10]. The split was carefully executed to maintain a balanced ratio of the output classes ("Yes" for weeks with one or more cases and "No" for weeks without any cases) in both the training and testing sets.

To further enhance the balance of the datasets, random oversampling was applied to equalize the number of records in each output class. This step was crucial in preventing the model from becoming biased towards the more frequent class, thereby improving its predictive capability.

Not only that, a comprehensive 3-fold cross-validation was employed to prevent overfitting and ensure the model's robustness [11]. In this method, the training data was divided into three equal subsets. The model was trained and validated three times, each time using a different subset as the validation set while the remaining two subsets (66.7% of the data) were used for training.

This process allowed the model to be tested on different segments of the data, providing a thorough evaluation of its performance. Each round of validation helped identify strengths and weaknesses, leading to better fine-tuning and optimization. By ensuring the model was exposed to varied data during training and validation, the 3-fold cross-validation enhanced the model's ability to generalize to new, unseen data, resulting in a more reliable and accurate predictive model.

2.7 Model training

Random forest [12], a machine learning technique that was also employed by [13], is a powerful ensemble method in machine learning that utilizes multiple decision trees created through the bagging (Bootstrap Aggregation) technique. This technique involves generating multiple resampled datasets from the original training set, where each dataset contains duplicate records produced by random sampling with replacement. This approach helps in diversifying the trees and reducing overfitting.

Each decision tree in the random forest is built independently on a resampled dataset. At each node of the tree, a subset of independent variables is randomly selected. The node is then split using the best variable from this subset, determined by a criterion like Gini impurity, which measures the probability of misclassification when splitting features at the nodes. This splitting process continues recursively until a specified stopping criterion is met, such as reaching a minimum node size or maximum tree depth.

After constructing all decision trees in a random forest model, predictions are aggregated for each input instance. In classification tasks, the final prediction is determined through majority voting across all trees. This means that for each input, the prediction that occurs most frequently among all the decision trees is selected as the final output.

In practice, a random forest model undergoes an internal model fitting procedure within the training set. This process is often performed multiple times, such as the 3-fold cross-validation that the study had employed, where the training data is split into three subsets. Each iteration uses two-thirds of the total training data for training and the remaining one-third for validation. This approach helps in evaluating and refining the model's performance by testing it on different subsets of the data.

Fig. 4 illustrates the sequential steps involved in the random forest algorithm. Beginning with the creation of resampled datasets through bagging, each decision tree is constructed independently on these datasets. The nodes of each tree are split based on randomly selected variables, aiming to optimize the predictive accuracy through diverse tree structures. By aggregating predictions from all trees, random forest effectively harnesses the collective strength of individual trees, making it a versatile and widely used method in machine learning for various predictive tasks.

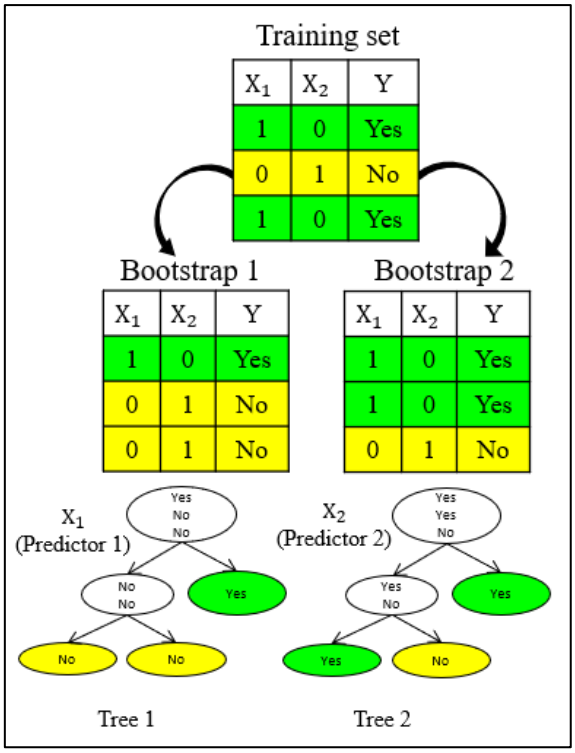

Fig. 4. Formation of random forest.

2.8 Model testing

The testing phase of the study was crucial for evaluating the model's predictive capability and robustness. To achieve this, 40% of the data from each 5 km by 5 km grid cell was designated as testing sets. These testing sets were essential for assessing the model's ability to forecast the presence or absence of leptospirosis cases on a weekly basis. During the testing phase, the single trained model was applied to each grid cell individually. For each 5 km by 5 km pixel, the model predicted which weeks would have reported cases of leptospirosis and which weeks would not. This process was repeated for all the testing sets across the study area, allowing for a thorough evaluation of the model's performance.

2.9 Model performance measurement

The developed models underwent evaluation using metrics such as testing set accuracy, sensitivity, and specificity [14]. Typically, accuracy alone might suffice to assess the overall predictive capability of the model. However, given the imbalance in output classes,

sensitivity and specificity metrics were also calculated to provide additional insights into the model's performance. The formulae for calculating these metrics are as follows:

Accuracy,

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity,

$$\frac{TP}{TP + FN}$$

Specificity,

$$\frac{TN}{TN + FP}$$

where,
TP = True Positives [correctly predicted positive (Yes) cases]
TN = True Negatives [correctly predicted negative (No) cases]
FP = False Positives [incorrectly predicted as positive (Yes)]
FN = False Negatives [incorrectly predicted as negative (No)]

According to the equations above, accuracy provides an overall measure of how often the model correctly predicts both positive (Yes) and negative (No) instances. It is influenced by the balance of the dataset classes; for balanced datasets, accuracy can provide a reliable measure of performance. Sensitivity indicates the proportion of actual positive instances that the model correctly identifies as positive (Yes). It is crucial in applications where correctly identifying positive cases is prioritized, such as in medical diagnostics or anomaly detection. Specificity indicates the proportion of actual negative instances that the model correctly identifies as negative (No). It complements sensitivity by focusing on the model's ability to avoid false positives, which is important in scenarios where correctly identifying negatives is critical, such as in spam email detection or quality control.

2.10 Receiver operating characteristics (ROC)

The Receiver Operating Characteristic (ROC) [15] curve was a critical tool for evaluating the performance of the leptospirosis prediction model. By plotting the true positive rate (sensitivity) against the false positive rate (1-specificity), the ROC curve provided a comprehensive view of the model's diagnostic ability across various threshold settings. This graphical representation helped in assessing the accuracy and robustness of the model's predictions.

To generate the ROC curve, the trained model was applied to the 40% testing sets of each 5 km by 5 km grid cell to predict weeks with and without leptospirosis cases. For each threshold value, the model's predictions were compared with the actual occurrences, calculating the true positive and false positive rates. The resulting ROC curve illustrated the trade-off between sensitivity and specificity, with the area under the curve (AUC) serving as a single scalar value summarizing the model's overall performance. A high AUC value indicated strong discriminative ability, showing that the model effectively distinguished between weeks with and without leptospirosis cases.

Based on Fig. 5., when the ROC curve extends above the diagonal line, it signifies that the model is effectively distinguishing between positive and negative cases better than random guessing. This indicates that the model exhibits discrimination ability beyond chance. Moreover, as the ROC curve ascends higher above the diagonal line towards the top-left corner, it reflects the model's enhanced capability to accurately classify positive and negative instances across varying threshold levels.

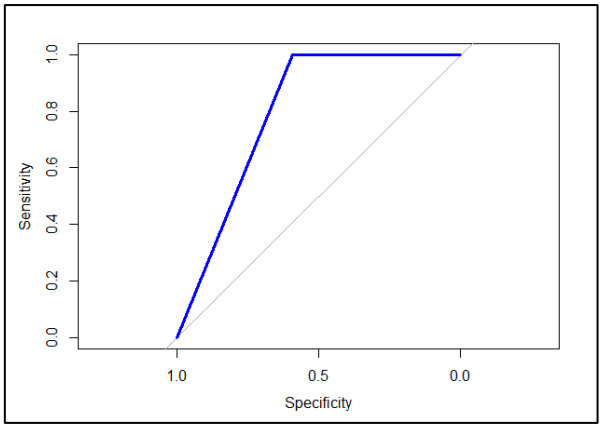

Fig. 5. ROC curve extending above diagonal line.

When the ROC curve falls below the diagonal line, as illustrated in Fig. 6., it suggests that the model's ability to distinguish between positive and negative cases is worse than random guessing. This indicates a lack of discrimination ability in the model's predictions. As the curve descends further below the diagonal line towards the bottom-right corner, it signifies a diminishing capability of the model to correctly classify positive and negative instances across different thresholds.

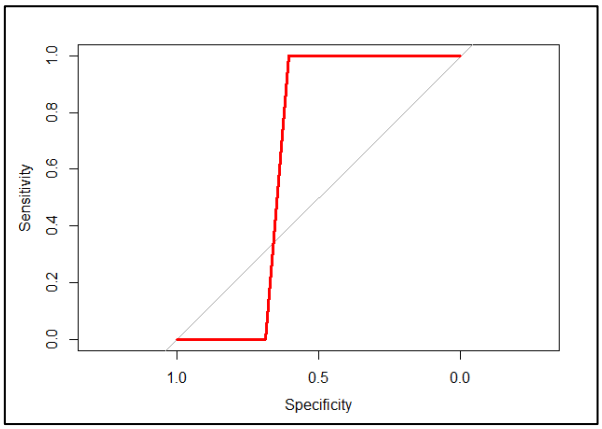

Fig. 6. ROC falling below the diagonal line.

2.11 Model scenarios

Four different model scenarios were tested, each incorporating different lag times, which are known to be crucial in understanding the hydrometeorological influences on leptospirosis [16]. The lag times were applied by shifting hydrometeorological data forward to align with the occurrence of cases, with intervals of 4 weeks ranging from 0 to 12 weeks in each model scenario.

3. RESULTS AND DISCUSSION

In total, four random forest models that differ by lag times were trained and tested. Table 2 presents the

performance of model during the training stage while Table 3 presents performance of model (after maximizing based on ROC) during the testing stage.

Table 2. Model performance metrics during training period.

Model Scenario	Accuracy (%)	Sensitivity (%)	Specificity (%)
0 week lag	98.2	100.0	96.3
4 weeks lag	88.7	85.2	83.6
8 weeks lag	90.6	84.5	86.7
12 weeks lag	91.1	85.0	87.5

Based on Table 2, the models demonstrate strong performance metrics: an average training accuracy of 92.2%, sensitivity of 88.7%, and specificity of 88.5%. These figures indicate that the models effectively learn and distinguish between weeks with and without leptospirosis cases on a weekly basis across a spatial resolution of 5km by 5km in the Kuantan district. The average sensitivity of 88.7% highlights the models' ability to accurately predict weeks when leptospirosis cases occur, demonstrating their effectiveness in both temporal and spatial dimensions. This capability ensures reliable identification of periods with actual leptospirosis cases within the specified geographic area. In terms of specificity, averaging at 88.5%, the models excel in predicting weeks without leptospirosis cases, ensuring dependable identification of periods when no cases were present within the same spatial resolution.

Furthermore, a minimal variation observed among different model scenarios underscores their stability and reliability across various time lag settings. This stability suggests that adjusting lag time may not significantly impact the models' performance, which is advantageous for saving training time and optimizing resource efficiency without compromising predictive accuracy.

Overall, these findings underscore the robustness of the random forest models in predicting leptospirosis, affirming their suitability for practical application in public health management and decision-making processes.

Table 3. Model performance metrics during testing period.

Model Scenario	Accuracy (%)	Sensitivity (%)	Specificity (%)
0 week lag	74.0	80.7	67.3
4 weeks lag	72.1	85.2	59.0
8 weeks lag	73.1	84.5	61.8
12 weeks lag	74.3	85.0	63.6

According to Table 3, The average testing accuracy of 73.4%, sensitivity of 83.9%, and specificity of 62.9% indicate a moderate performance of the model in predicting leptospirosis cases. While the overall accuracy remains decent, the model demonstrates a strong sensitivity in identifying weeks with leptospirosis cases, which is crucial for early detection and intervention by public health officers. However, the lower specificity suggests that the model is less accurate in correctly identifying weeks without leptospirosis cases, potentially leading to more false positives.

A notable observation is the discrepancy between training and testing performance. The model appears to perform better during the training period than testing period, indicating a potential issue of overfitting. Overfitting occurs when the model fits too closely to the training data and fails to generalize well to new, unseen data.

To address this, it may be necessary to increase the number of repeats or iterations during the internal model fitting process, such as cross-validation. By performing more iterations of cross-validation, the model can be exposed to a greater variety of training subsets and validation folds. This helps in generating a more generalized model that better captures underlying patterns and relationships in the data, rather than memorizing specific instances from the training set.

Enhancing the model's generalization ability through increased cross-validation iterations can mitigate overfitting, improving its performance on unseen testing data. This approach ensures that the random forest model maintains robustness and reliability in real-world applications, supporting more accurate and effective decision-making in public health management.

4. CONCLUDING REMARKS

The models achieved an average prediction accuracy of 73.4% in anticipating the occurrence of leptospirosis cases within the Kuantan district, establishing them as a valuable tool for public health officers. This accuracy level allows officers to effectively discern weeks with and without leptospirosis cases, facilitating timely interventions and resource allocation. However, these findings are preliminary, and there is potential to enhance model performance through meticulous adjustment of model controls and parameters.

A noteworthy discovery from this study is the minimal variability observed across different model scenarios. This suggests that adjustments in lag time may have limited impact on the model's predictive capabilities when employing random forest machine learning for leptospirosis prediction. This stability in performance underscores the reliability of the models in various temporal configurations, offering consistency in their ability to forecast disease occurrence.

Moving forward, further refinement and optimization of model parameters could potentially improve accuracy and robustness. By fine-tuning these aspects, future iterations of the models could yield even more precise predictions, bolstering their utility as essential tools in public health decision-making and proactive management strategies.

5. REFERENCES

[1] Levett, P. N. (2001). Leptospirosis. Clinical Microbiology Reviews, 14(2), 296–326. https://doi.org/10.1128/CMR.14.2.296

[2] Karpagam KB, Ganesh B. Leptospirosis: a neglected tropical zoonotic infection of public health

importance-an updated review. Eur J Clin Microbiol Infect Dis. 2020 May;39(5):835-846. doi: 10.1007/s10096-019-03797-4. Epub 2020 Jan 2. PMID: 31898795.

[3] Guerra, M. A. (2013). Leptospirosis: Public health perspectives. Biologicals, 41(5), 295–297. https://doi.org/10.1016/j.biologicals.2013.06.010

[4] Dhewantara, P. W., Lau, C. L., Allan, K. J., Hu, W., Zhang, W., Mamun, A. A., & Soares Magalhães, R. J. (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

[5] Edre, M. A., Hayati, K. S., Salmiah, M. S., & SI, S. N. (2015). A case control study on factors associated with leptospirosis infection among residents in flood-prone area, Kuantan: a geographical information system-based approach. International Journal of Public Health and Clinical Sciences, 2(3), 151-163.

[6] Oliver, M. A., & Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. International Journal of Geographical Information System, 4(3), 313-332.

[7] Hamisi, J. (2013). Study of rainfall trends and variability over Tanzania (Doctoral dissertation).

[8] Imdadullah, M., Aslam, M., & Altaf, S. (2016). mctest: An R Package for Detection of Collinearity among Regressors. R J., 8(2), 495.

[9] Joenel, G., & Jonathan, D. (2023, October). Estimation of Permeability of Soil-Fly Ash Mix using Machine Learning Algorithms. In *Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES)* (Vol. 9, pp. 28-33). Interdisciplinary Graduate School of Engineering Sciences, Kyushu University.

[10] Ucar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The Effect of Training and Testing Processon Machine Learning in Biomedical Datasets. Mathematical Problems in Engineering, 2020. https://downloads.hindawi.com/journals/mpe/2020/2836236.pdf

[11] Berrar D. (2018) Cross-validation. Encyclopedia of Bioinformatics and Computational Biology, Volume 1, Elsevier, pp. 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

[12] Breiman, L. (2001). Random Forests. Machine learning, 45(1), 5-32. https://doi.org/10.1007/978-3-030-62008-0_35

[13] Luzorata, J. G., Bocobo, A. E., Detera, L. M., Pocong, N. J. B., & Sajonia, A. P. (2023, October). Assessment of Land Use Land Cover Classification using Support Vector Machine and Random Forest Techniques in the Agusan River Basin through Geospatial Techniques. In *Proceedings of International Exchange and Innovation Conference on Engineering & Sciences (IEICES)* (Vol. 9, pp. 240-246). Interdisciplinary Graduate School of Engineering Sciences, Kyushu University.

[14] Van Stralen, K. J., Stel, V. S., Reitsma, J. B., Dekker, F. W., Zoccali, C., & Jager, K. J. (2009). Diagnostic methods I: sensitivity, specificity, and other measures of accuracy. Kidney international, 75(12), 1257-1263.

[15] Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine, 8(1), 19-20.

[16] Rahmat, F., Ishak, A. J., Zulkafli, Z., Yahaya, H., & Masrani, A. (2019). Prediction model of leptospirosis occurrence for Seremban (Malaysia) using meteorological data. International Journal of Integrated Engineering, 11(4), 61–69. https://doi.org/10.30880/ijie.2019.11.04.007