INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION
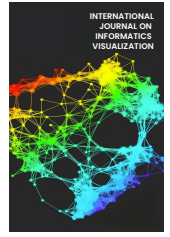
Predicting Dengue Outbreak based on Meteorological Data Using Artificial Neural Network and Decision Tree Models

Nor Farisha Muhamad Krishnan [a,*], Zuriani Ahmad Zukarnain [b], Azlin Ahmad [a], Marhainis Jamaludin [b]

[a] Faculty Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[b] Faculty Computer and Mathematical Sciences, Universiti Teknologi MARA, Bukit Ilmu, 18500 Machang, Kelantan, Malaysia
Corresponding author: *farisha.krishnan@gmail.com

Abstract— Dengue fever is well-known as a potentially fatal disease, and the number of cases in some areas remains uncontrolled. Despite efforts to prevent the dengue outbreak from spreading further, vectors may be to blame. Identifying what weather characteristics contribute to dengue outbreaks is important to predict the dengue outbreak. This study proposes Artificial Neural Network (ANN) and Decision Tree (DT) models based on maximum temperature, minimum temperature, total rainfall, and average humidity to predict the dengue outbreak in Kota Bharu. Different numbers of hidden nodes were used in ANN to optimize the model. Both models, ANN and DT are evaluated based on accuracy, sensitivity and specificity showing that ANN (Accuracy = 68.85%, Sensitivity = 99.71%, Specificity = 1.27%), performed better than DT (Accuracy = 67.46%, Sensitivity = 98.82%, Specificity = 2.53%). This means that ANN outperforms DT when predicting a dengue outbreak in Kota Bharu. Based on the ANN model, it can be concluded that the number of hidden nodes affects the model's accuracy. Selecting the ideal number of hidden nodes for modeling the ANN model is appropriate. Even though ANN accuracy for prediction models is greater than DT, it is still low. It can be inferred that selecting a prediction model appropriate for a variety of dataset types and levels of complexity is important. Based on these models, the government may take pre-emptive actions to enhance public awareness about climate change.

Keywords— Dengue; climate; Kota Bharu; artificial neural network; decision tree.

I. INTRODUCTION

Dengue fever is a mosquito-borne disease that causes a fever, severe headache, rashes, muscle, and joint pain, pain behind the eyes, and, in rare cases, bleeding. This disease may have a role in the rise in morbidity and death in Malaysia and elsewhere. Dengue fever is a mosquito-borne viral infection that has become one of the world's most quickly spreading and dangerous diseases [1]. They defined the dengue outbreak response as the collection of actions to lower fatality rates, case counts, and other aspects of a dengue outbreak. In Malaysia, a dengue outbreak happens when two or more dengue cases are recounted from the same location within seven days [2], [3].

According to the World Health Organization [1], in a certain community, geographical area, or period, an epidemic is described as an increase in the number of cases of disease beyond what was expected. Human behavior, climate, and insects are the vectors that have produced the dengue outbreak. The major vector is the mosquito Aedes aegypti, which is

highly adapted to the human environment. The vectors are quite susceptible to variations in the weather pattern since they are primarily present during the rainy season when rainfall helps to create vector-breeding habitats [4]. According to Ochida [5], dengue dynamics result from intricate interactions between the host, the virus, and the vector that are influenced by the environment. Several studies studied the relationships between dengue dynamics and weather and the impact of climate variation on these dynamics. Furthermore, controlling the Dengue fever outbreak in tropical nations is a major public health concern. Precision prediction approaches are in high demand since effective vector-control activities are contingent on the timeliness of public initiatives [6]. Machine learning technology is being utilized to better anticipate the epidemic [7]. The vector control program aims to prevent dengue transmission and is extensively utilized, although it is not always successful. The explanation might be that the poor and government focus on epidemic interventions too late [4]. As a result, predicting a dengue outbreak may aid in developing early warning systems,

which may subsequently be used to improve the effectiveness of the vector control program.

Dengue fever is well-known as a disease that may be fatal, and there are still certain regions where the number of cases is uncontrolled. Although more attempts have been made to prevent the dengue outbreak from spreading, it is possible that vectors are to blame. Depending on the environment, lifestyle, weather, and other factors, various areas may have different factors that increase dengue outbreaks. Several previous studies have focused on Peninsular Malaysia's west coast and have been location specific. Despite having the second-highest number of dengue cases on Peninsular Malaysia's east coast, Kelantan seems to have a little study on the illness [7]. There may be some; however, they are slightly old research, making it difficult to compare because the condition and circumstances may have changed since the last study as urbanization has progressed. Besides, the factors contributing to dengue outbreak might vary from region.

A. Literature Review

With the development of data mining technologies and the availability of usable data, important information in a variety of fields can be discovered. In data mining, machine learning algorithms are used to excerpt meaningful correlations and information from massive amounts of data and to give an automatic tool for various classifications and predictions [8]. Machine learning is a technique for predicting and classifying new datasets using a mathematical model to learn patterns from prior datasets [9]. Machine learning is also employed, according to Daw, since it can process data more efficiently. The Artificial Neural Network (ANN) and Decision Tree (DT) models could be used to predict the dengue epidemic in this study. Both models are used because they are easy to apply and require setting a few variables to deal with high-dimensional input. Health organizations use data mining to categorize diseases like dengue fever, diabetes, and cancer [10]. In order to build predictive models for predicting dengue severity in patients, machine learning techniques were proposed. The artificial neural network has the discriminative ability for severe dengue prediction outperformed support vector classifier, logistic regression, gradient boosting machine, and random forest. During dengue outbreaks, this method could help the related authorities to make a quick diagnosis [11].

Various machine learning approaches may be used to forecast dengue outbreaks. Ibrahim et al. [12] employed a clustering technique to forecast dengue fever-infected areas, and the method proved to be successful in identifying the infected region of the dengue epidemic in Malaysia. The regions are the busiest in Malaysia, having a large population that encourages the spread of dengue fever. The machine learning algorithms were developed based on non-clinical data and non-traditional sources to predict outbreaks and decrease reporting delays. The researchers wanted to find studies that tracked and predicted dengue-related outcomes using Big Data, real-world data, and/or machine learning techniques [13].

Machine learning offers much promise for forecasting dengue outbreaks [14]. To construct a dengue prediction model, future research should look towards using nature-inspired algorithms or boosting. Ughelli [15] conducted research to estimate the number of dengue cases and identify additional factors that influence the number of cases. He implemented the proposed model using ANN since the approach can discover patterns between non-linear input variables and variables to predict. Research is comparable to this one in that it employed the same variables as this one: rainfall, temperature, humidity, and previous dengue cases.

Predictions from untrained data should be verified by comprehensively reflecting accuracy, recall, precision, and other factors. Predictive models generated from training data should verify the accuracy of the predictions compared to the verification data, and predictions from untrained data should be verified by comprehensively reflecting not only accuracy but also recall, precision, and other factors [16].

McGough [17] built a novel machine learning dengue prediction system that detects local weather and population vulnerability patterns dynamically in time and place to generate epidemic prediction months ahead of outbreaks at the urban level in Brazil. Weather-based predictions increased when population susceptibility data was included, suggesting that immunity is a significant predictor that most dengue prediction models neglect. Kamarudin [18] evaluated various previous studies and concluded that a combination of important components and time-series forecasting analysis would deliver a foundation for constructing adequate predictive analytics frameworks. The method is seen to be necessary, and it could help link the gap between dengue cases, human responses, and vector population dynamics,

It can be difficult to predict the future using time series since many unknowns exist, such as natural catastrophes, legislation, and climate conditions [19]. Their study compared multilayer perceptron (MLP) artificial neural networks to the Root Mean Squared Error (RMSE) and coefficient of determination (R2) depending on Autoregressive Integrated Moving Average (ARIMA). MLP is better than ARIMA in forecasting the future since it has the highest coefficient of determination and the least prediction error. Balasaravanan and Prakash [20] investigated whether the average temperature, average humidity, rainy days per week, total rainfall, and past dengue cases influence current dengue cases using ANN and Classification And Regression Trees (CART). It was discovered that ANN outperforms CART in terms of accuracy, scoring 92 percent versus 88 percent, respectively. According to ANN, those characteristics are strongly connected with reported dengue cases at specific time intervals.

Machine learning was a technique for predicting and classifying new datasets by applying a mathematical model to learn patterns from prior datasets [9]. Machine learning was also deployed, according to Daw, because it can process data more efficiently. According to the study, the decision tree model was also used to identify the relevant factors for the prediction process. Decision Trees communicate inherent decision-making logic, which is not present in black box algorithms like Neural Networks [21]. It takes less time to train compared to the neural network approach. Another study by Ceballos-Arroyo et al. [22] used meteorological, epidemiological, and environmental data and modeled an artificial neural network as the foundation of a machine learning system. This study aimed to detect future Dengue outbreaks so that an early warning system could be developed,

a schedule could be established for health officials to create contingency measures, and the consequences of Dengue on Medellin's population could be reduced.

The neural networks were more accurate in predicting outbreak peak intensity, a yearly dengue time series, and peak timing. Furthermore, incorporating human motion data into the neural network might improve its accuracy [23]. According to Benedum et al. [24], predictive models can be utilized as primary warning systems to predict the risk of future pandemic diseases. Regression and time series models based on data from dengue observation (e.g., case counts) and weather are commonly used to forecast dengue fever. However, the number of predictors and model assumptions that can be incorporated into these models may be limited. Machine learning (ML) models, according to the study, are also a potential alternative because they have a non-linear structure and can be practical to data with many dimensions.

II. MATERIAL AND METHOD

A. Source of Data

This study acquired secondary data from Kelantan's Department of Health and Meteorological Department. The data covers all districts in Kelantan from 2015 to 2019 (total data around 10467) and contains date, districts, the region, identity number, and notification number. The total number of dengue cases registered in Kelantan, is based on the statistics provided by Kelantan's Vector Borne Disease Control. The climatic variables of Kota Bharu collected from the Meteorological Department (total data around 1827) include the date, maximum temperature, minimum temperature, total rainfall, and average humidity. To generate the target variable, the dengue outbreak is based on the quantity of dengue cases reported in a certain region within two weeks. The target would be a binary which dengue outbreak for yes or no. Briefly, the analysis in this study would only include data for maximum temperature, minimum temperature, total rainfall, and average humidity to predict the dengue outbreak. Both datasets would be merged to obtain the target variable based on the date recorded. Thus, the total data would reduce to obtain the target variable, and the number of cases would be classified into a group which less than 2 cases and more than 2 cases.

B. Analysis Method

Since this study was based on secondary data, pre-processing was required, such as data cleaning, eliminating duplicates, and dealing with missing variables. In general, duplicate data can be found in databases, which can cause findings to be skewed. To identify replication, unique identification is required. If the redundant data is eliminated, the performance would improve. The approaches for removing present duplication in data cleaning are at the heart of determining the similarity between consecutive documents in a well-organized database [25]. Missing values in datasets can cause severe problems for data mining and machine learning systems that are not prepared to handle them, such as knowledge loss, higher failures, and biased outcomes. The missing data, in actuality, is a source of vulnerability in learning processes, and it has the potential to degrade data quality [26], [27].

After that, continue to clean the dataset by examining the dataset's outliers. Outliers should be handled correctly to optimize model performance. Data points deviate considerably from others, or observations that are tiny or big concerning the vast majority of observations, are known as outliers [28], [29]. Outliers may reveal much about a topic and the data collection method. Box plots, histograms, interquartile ranges, scatterplots, and z-scores are some methods for detecting outliers [30]. Certain techniques would be used to address noisy data, outliers, and missing and duplicate data. The data is then converted into a different format to portray it per the study's goals. Because the data does not have the same scale, each column would be normalized to enhance efficiency. The data were standardized since the independent variables' values fluctuated greatly and had different units [31].

Predictive models are then employed in the modeling stage to predict the dengue epidemic. If more than one technique is used, the work must be completed independently for each technique. A process to evaluate the model's quality is required before it can be built. The data would be allocated into training and testing datasets before modeling begins. In machine learning classification, imbalanced groups, in which each class contains a disproportionate number of observations, are a typical issue. Dealing with imbalanced data when conducting classification on datasets is a common problem. The data would be separated into training (70%), and testing (30%) data sets to perform a model on an out-of-sample dataset. The validation set is not utilized here to maintain the focus on the neural network and decision tree model.

The dengue and climate datasets are utilized to forecast and classify algorithms and compare their performance using R. Sensitivity, specificity, and accuracy are the measures used to assess the algorithms' performance. A test's accuracy is defined as its ability to distinguish between healthy and patient instances accurately, its sensitivity as its ability to identify patient instances correctly, and its specificity as its ability to locate healthy examples [32] correctly. To evaluate the model's efficiency, the metrics can be represented as a confusion matrix, also known as a contingency table. It shows the model's overall accuracy, and the proportion of the classifier successfully recognized total samples. This study has two classifications for dengue outbreak output: yes and no.

Maximum temperature, minimum temperature, total rainfall, and average humidity are the variables for the input layer because the data has already been standardized. The following step was linked to the hidden layer. Because no precise theory exists that specifies the optimum approach for determining the hidden nodes, the study tried to change from 1 to 7 hidden nodes, as there was only four inputs. Based on the rules of thumb, the number of hidden neurons should be twice the number of input [33]. Because the model's output was binary, this study examined the number of hidden neurons employed and the model's accuracy. For binary classification, the MSE function is non-convex. The most accurate model would be utilized to build a model based on the number of hidden neurons.

Since the dependent variable is a category variable (binary) and the predictor (independent) variables can be continuous, this study used the Classification and Regression Tree (CART). The decision tree was the next model chosen for

analysis since it is a good and practical approach that can be easily converted into basic classification rules and can be constructed very rapidly. The procedure begins with the selection of the variable that yields the best split (based on the lowest Gini Index). As this study used a binary dependent variable, the highest Gini index value is 0.5 (the dengue outbreak: yes or no). Since Gini does not require a log calculation, it is faster than Entropy. The result was correctly classified if the gradient was darker, while the output was slightly incorrectly classified for the dengue outbreak if the gradient was lighter.

Finally, the classification table is utilized to complete the accuracies of the models, which is done with the test dataset. From the confusion matrix, the accuracy can be calculated. The accuracy of the Artificial Neural Network and the Decision Tree model in predicting dengue outbreaks were compared using the confusion matrix result. The correctly categorized dengue epidemic may be recognized using the classification table (see Table 1), which has correctly classified as yes or no. The models were assessed using the classification table's accuracy, sensitivity, and specificity.

TABLE I
CLASSIFICATION TALE

		Actual	
		Positive	**Negative**
Predicted	Positive	TP	FP
	Negative	FN	TN

III. RESULT AND DISCUSSION

As inputs for this study were 4, and the hidden neuron should be less than twice the input based on the rules of thumb [33], this study examined the error with different numbers of hidden nodes ranging from 1 to 7. The accuracy of the model was compared using different numbers of hidden neurons in the study. The model with two hidden neurons was utilized for the ANN model since it offers the best results. The number of hidden nodes employed in ANN has been found to affect its performance.

TABLE II
THE ACCURACY OF VARYING NUMBER OF HIDDEN NEURONS

Number of Hidden Neurons	Accuracy (%)
1	68.65
2	68.85
3	68.65
4	68.25
5	67.86
6	67.26
7	67.46

The black lines represented the connections between each layer and the weights on each link, according to Alice [34], while the blue lines represented the bias term applied in each step. The ANN model with four input parameters (MaxTemp, AvgHum, MinTemp, and Rainfall) and binary output is shown in Fig. 1.

Accuracy, specificity, and sensitivity were calculated based on the classification table of this study. This model's confusion matrix is presented in Table III.
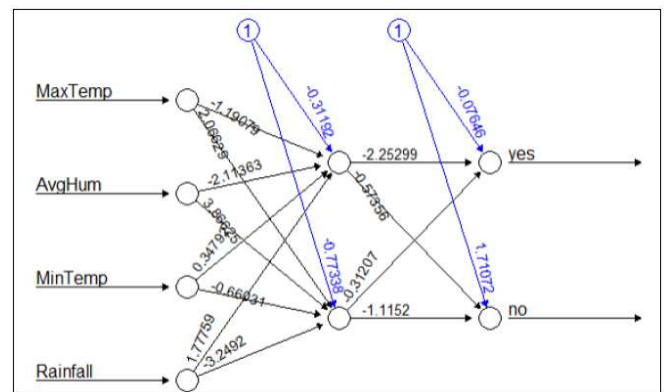

Fig. 1 Artificial Neural Network Model

TABLE III
CLASSIFICATION TABLE FOR ARTIFICIAL NEURAL NETWORK MODEL

		Predicted	
		Yes	**No**
Actual	Yes	345	156
	No	1	2

The ANN model in Table III correctly predicted 345 records of the dengue outbreak and 2 records of no dengue outbreak. Meanwhile, 156 dengue outbreak records were incorrectly predicted as no dengue outbreaks, while only 1 record for no dengue outbreak was incorrectly predicted as a dengue outbreak. The performance of the ANN model was then assessed using ANN classification table.

$$Accuracy = \frac{345 + 2}{345 + 156 + 1 + 2} = 0.6885 \quad (1)$$

$$Sensitivity = \frac{345}{345 + 1} = 0.9971 \quad (2)$$

$$Specificity = \frac{2}{2 + 156} = 0.01266 \quad (3)$$

The decision tree was the next model chosen for analysis since it is a good and practical approach that can be easily converted into basic classification rules and can be created very rapidly. Plotting the decision tree model in R is using package rpart. The function `rpart()` is used to build the model and `rpart.plot()` is used to plot the decision tree model. The dengue outbreak alert's overall likelihood is presented at each tree's end. This study employed the CART (Classification and Regression Tree) model. To begin, create a split test and score each for the possible splits using a goodness measure score, such as the Gini Index in this study.

TABLE IV
GINI INDEX VALUES FOR EACH VARIABLE

Variables	Gini Index
MaxTemp	0.1846409
MinTemp	0.1991777
AvgHum	0.2317518
Rainfall	0.8135272

MaxTemp had the lowest Gini Index value; hence the model would begin splitting based on MaxTemp. Since the Gini Index of rainfall is greater than 0.5, the parameter was

excluded from the model. Look at Fig. 2, a plot for the decision tree model. The chosen tree has three splits.
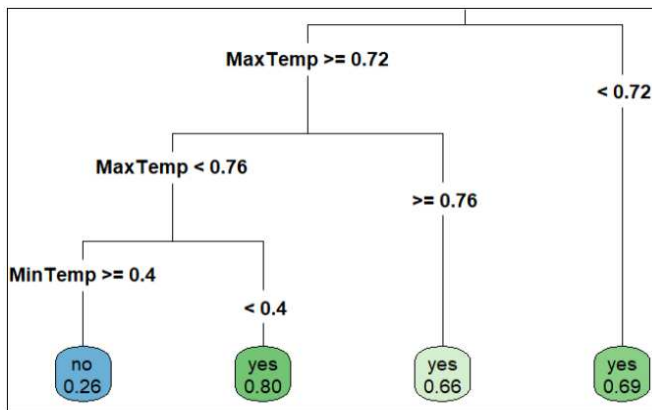

Fig. 2 Decision Tree Model

The bottom level node's color gradient suggests that a dengue outbreak with a background arrangement from the branch is expected to be either yes or no. The more vibrant the color, the more likely it is to appear in the labeled output. As the number of splits grows, the complexity of DTs grows. In general, simpler DTs are preferred over more complicated ones since they are better to understand and less prone to overfitting. However, because only four inputs were used in this study, the decision tree model was not overly complex. It did not have a lot of splits, so it did not need to be pruned. From the confusion matrix, the accuracy can be calculated. The confusion matrix result for the decision tree model is shown in TABLE V.

TABLE V
CLASSIFICATION TABLE OF THE DECISION TREE MODEL

		Predicted	
		Yes	No
Actual	Yes	336	154
	No	10	4

The ANN model correctly predicted the 336 dengue outbreak records based on the DT model classification table in `Table` V, but only 4 records of no dengue outbreak. Meanwhile, 156 dengue outbreak data were misidentified as no dengue outbreaks, and 10 no dengue outbreak records were misidentified as dengue outbreaks. The DT model, like ANN, would be assessed based on the classification table value.

$$Accuracy = \frac{4+336}{4+154+10+336} = 0.6746 \qquad (4)$$

$$Sensitivity = \frac{336}{336+10} = 0.9882 \qquad (5)$$

$$Specificity = \frac{4}{4+154} = 0.0253 \qquad (6)$$

Table 6 shows that ANN outperforms DT in terms of accuracy, specificity, and sensitivity in predicting dengue outbreaks. Fig. 3 demonstrates that ANN has 68.85% higher accuracy than DT, which has 67.46% accuracy. In terms of sensitivity, ANN outperforms DT by 99.71% and 98.82%, respectively, as shown in Fig. 4. The ability of the test to correctly detect dengue epidemics is referred to as sensitivity.

However, DT has a higher specificity than ANN, with 2.53% and 1.27%, respectively, as shown in Fig. 5.

TABLE VI
COMPARISON BETWEEN ANN AND DT

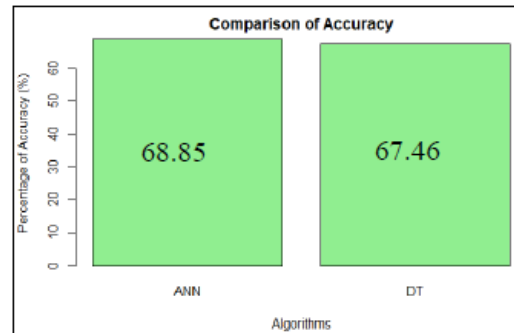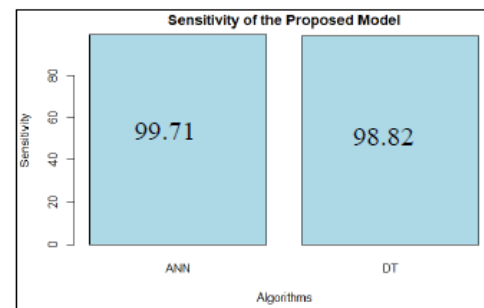|	ANN	DT
Accuracy	68.85%	67.46%
Sensitivity	99.71%	98.82%
Specificity	1.27%	2.53%

Fig. 3 Comparison of Accuracy for ANN and DT

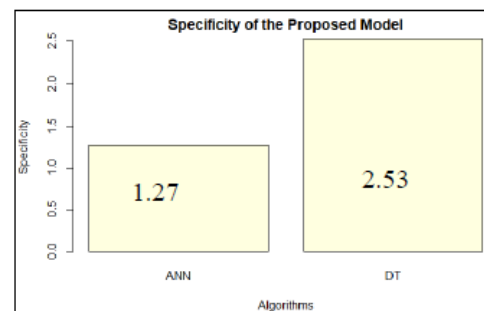Fig. 4 Sensitivity of ANN and DT

Fig. 5 Specificity of ANN and DT

This study aims to predict dengue outbreaks using weather parameters, including average humidity, maximum temperature, minimum temperature, and rainfall. After analysis, it can be found that all factors are affecting the dengue outbreak when using an artificial neural network, while in the decision tree, only rainfall does not include in the modeling decision tree for the outbreak of dengue prediction. When accuracy was utilized to evaluate the performance of a decision tree (DT) and an artificial neural network, the artificial neural network (ANN) outperformed the decision tree (DT). It was discovered that ANN scored 68.85 % over DT (67.46 %). This indicates that when predicting a dengue epidemic in Kota Bharu, ANN outperforms DT. According to

the model prediction for ANN, the number of hidden nodes used impacts the model's accuracy.

In this experiment, neural networks outperformed decision tree classifiers as they can identify useful information that extracts hidden patterns from huge, complicated, and inaccurate data and find data trends humans and other computational methods cannot easily perceive. As it can evaluate both numerical and categorical data inputs, the decision tree strategy performs almost as well as the neural network technique in this experiment. [35]. It can also handle enormous datasets, and its classifier can quickly recognize and comprehend relationships between predictor variables, particularly categorical data.

Both models have a poor level of accuracy. Normally, models generated using ANN have shown excellent accuracy [31]. However, the accuracy in this work is lower than normal. This may be due to the data's condition, which is that it is from a short time period and only covers one location, Kota Bharu, where the value is nearly the same. The data would be sorted into week and region, resulting in a reduction in the number of rows, in order to obtain the target variable (dengue outbreak) because the dengue outbreak defined 2 or more dengue cases recorded in 7 days in a specific region. The poor performance could be attributed to the model's limited number of inputs and the dataset's short duration. The study's average humidity, maximum temperature, minimum temperature, and rainfall were the only four inputs.

IV. CONCLUSION

As a result, the neural networks and decision tree classifiers do not perform well enough in classifying dengue outbreak predictions, with the accuracy of 68.85% and 67.46% for ANN and DT, respectively. Even if ANN accuracy is better than DT, the prediction model's accuracy is rather poor. Based on the ANN model, it can be deduced that the number of hidden nodes affects the model's accuracy. As a result, selecting the optimal number of hidden nodes for modeling the ANN model is appropriate. In summary, both the ANN and DT models have poor accuracy, but their sensitivity, or the fraction of true positives accurately categorized, is quite high, at 99.71% and 98.82%, respectively.

In the ANN model, all meteorological elements contribute to the dengue epidemic. However, only the rainfall parameter is not included in the DT model for dengue outbreak prediction. Based on these models, the government and the VBDC may take preventative measures to raise public awareness about climate change. Understanding climate behavior and implementing education initiatives can enhance the region's early warning system. Future research should focus on the seasons or periods when dengue fever is most prevalent and is expected to be more comprehensive, with more relevant climatic features, dengue data, and demographic factors.

REFERENCES

[1] World Health Organisation, "Dengue and severe dengue," *WHO Fact Sheet*, 2014, doi: 10.1111/1469-0691.12442.

[2] A. S. Masrani, N. R. Nik Husain, K. I. Musa, and A. S. Yasin, "Prediction of Dengue Incidence in the Northeast Malaysia Based on Weather Data Using the Generalized Additive Model," *Biomed Res. Int.*, vol. 2021, 2021, doi: 10.1155/2021/3540964.

[3] World Health Organisation (WHO), *Technical handbook for dengue surveillance, dengue outbreak prediction/detection and outbreak response ("model contingency plan")*. Geneva: Runge-Ranzinger, 2016.

[4] A. L. Ramadona, L. Lazuardi, Y. L. Hii, Å. Holmner, H. Kusnanto, and J. Rocklöv, "Prediction of dengue outbreaks based on disease surveillance and meteorological data," *PLoS One*, vol. 11, no. 3, pp. 1–18, 2016, doi: 10.1371/journal.pone.0152688.

[5] N. Ochida *et al.*, "Modeling present and future climate risk of dengue outbreak, a case study in New Caledonia," *Environmental Health*, vol. 21, no. 1. 2022, doi: 10.1186/s12940-022-00829-z.

[6] C. Souza, P. Maia, L. M. Stolerman, V. Rolla, and L. Velho, "Predicting dengue outbreaks in Brazil with manifold learning on climate data," *Expert Syst. Appl.*, vol. 192, p. 116324, Apr. 2022, doi: 10.1016/J.ESWA.2021.116324.

[7] iDengue, "iDengue untuk Komuniti," 2020.

[8] E. Bahmani, M. Jamshidi, and A. A. Shaltooki, "Breast cancer prediction using a hybrid data mining model," *Int. J. Informatics Vis.*, vol. 3, no. 4, pp. 327–331, 2019, doi: 10.30630/joiv.3.4.240.

[9] S. Daw, "Machine Learning Applications Using Waikato Environment for Knowledge Analysis," no. Iccmc, pp. 346–351, 2020.

[10] A. H. S. B. Abu Hassan Shaari *et al.*, "A data mining approach for precise diagnosis of dengue fever," *Int. J. Recent Technol. Eng.*, vol. 9, no. 1, pp. 1–9, 2016, doi: 10.1038/nrmicro3430.

[11] S. W. Huang, H. P. Tsai, S. J. Hung, W. C. Ko, and J. R. Wang, "Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning," *PLoS Neglected Tropical Diseases*, vol. 14, no. 12. pp. 1–19, 2020, doi: 10.1371/journal.pntd.0008960.

[12] N. Ibrahim, N. S. Akhir, and F. H. Hassan, "Using Clustering and Predictive Analysis of Infected Area on Dengue Outbreaks in Malaysia," vol. 9, no. 2, pp. 51–58, 2018.

[13] E. Sylvestre *et al.*, "Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review," *PLoS neglected tropical diseases*, vol. 16, no. 1. p. e0010056, 2022, doi: 10.1371/journal.pntd.0010056.

[14] N. Azam *et al.*, "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques," *Sci. Rep.*, pp. 1–9, 2021, doi: 10.1038/s41598-020-79193-2.

[15] V. Ughelli, Y. Lisnichuk, J. Paciello, and J. Pane, "Prediction of Dengue Cases in Paraguay Using Artificial Neural Networks," *3rd Int'l Conf Heal. Informatics Med. Syst.*, pp. 130–136, 2017.

[16] H. S. Park and S. J. Yoo, "Early Dropout Prediction in Online Learning of University using Machine Learning," *JOIV Int. J. Informatics Vis.*, vol. 5, no. 4, p. 347, 2021, doi: 10.30630/joiv.5.4.732.

[17] S. F. McGough, L. Clemente, J. N. Kutz, and M. Santillana, "A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles," *Journal of the Royal Society Interface*, vol. 18, no. 179. 2021, doi: 10.1098/rsif.2020.1006.

[18] A. N. A. Kamarudin, Z. Zainol, and N. F. A. Kassim, "Forecasting the dengue outbreak using machine learning algorithm: A review," *2021 Int. Conf. Women Data Sci. Taif Univ. WiDSTaif 2021*, pp. 26–30, 2021, doi: 10.1109/WIDSTAIF52235.2021.9430193.

[19] A. Olawoyin, Y. Chen, A. Olawoyin, and Y. Chen, "ScienceDirect ScienceDirect ScienceDirect Predicting the Future with Artificial Neural Network Predicting the Future with Artificial Neural Network," *Procedia Comput. Sci.*, vol. 140, pp. 383–392, 2018, doi: 10.1016/j.procs.2018.10.300.

[20] K. Balasaravanan and M. Prakash, "Detection of dengue disease using artificial neural network based classification techniquetion," *Int. J. Eng. Technol.*, vol. 7, no. 1, pp. 13–15, 2018, doi: 10.14419/ijet.v7i1.3.8978.

[21] A. Navlani, "Decision Tree Classification in Python," *Data Camp*, 2018. .

[22] A. M. Ceballos-Arroyo *et al.*, "Towards a machine learning-based approach to forecasting Dengue virus outbreaks in Colombian cities: a case-study: Medellin, Antioquia," 2019.

[23] R. Bomfim *et al.*, "Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas: Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas," *Journal of the Royal Society Interface*, vol. 17, no. 171. 2020, doi: 10.1098/rsif.2020.0691.

[24] C. M. Benedum, K. M. Shea, H. E. Jenkins, L. Y. Kim, and N. Markuzon, "Weekly dengue forecasts in Iquitos, Peru; San Juan, Puerto Rico; and Singapore," *PLoS Neglected Tropical Diseases*, vol. 14, no. 10. pp. 1–26, 2020, doi: 10.1371/journal.pntd.0008710.

[25] W. L. Low, M. L. Lee, and T. W. Ling, "A knowledge-based approach for duplicate elimination in data cleaning," *Inf. Syst.*, 2001.

[26] S. M. Abu-soud, "A Novel Approach for Dealing with Missing Values in Machine Learning Datasets with Discrete Values," in *International Conference on Computer and Information Sciences (2019 ICCIS)*, 2019, no. July.

[27] Y. Dong and C. J. Peng, "Principled missing data methods for researchers," vol. 2004, pp. 1–17, 2013.

[28] H. Aguinis, R. K. Gottfredson, and H. Joo, "Best-Practice Recommendations for Defining, Identifying, and Handling Outliers," *Organ. Res. Methods*, vol. 16, no. 2, pp. 270–301, 2013, doi: 10.1177/1094428112470848.

[29] D. Cousineau and S. Chartier, "Outliers detection and treatment : A review Outliers detection and treatment : a review .," *Int. J. Psychol. Res.*, vol. 3, no. 1, pp. 59–68, 2010, doi: 10.21500/20112084.844.

[30] J. Brownlee, "How to Remove Outliers for Machine Learning," *Machine Learning Mastery*, 2018.

[31] P. Silitonga, B. E. Dewi, A. Bustamam, and H. S. Al-Ash, "Evaluation of Dengue Model Performances Developed Using Artificial Neural Network and Random Forest Classifiers," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 135–143, 2021, doi: 10.1016/j.procs.2020.12.018.

[32] A. Baratloo and S. Safari, "Evidence based medicine, simple definition and calculation of accuracy, sensitivity and specificity," *Iran. J. Emerg. Med.*, vol. 2, pp. 105–107, 2015.

[33] J. Heaton, *Introduction to Neural Networks with Java*, Second Edi. Heaton Research, Inc, 2008.

[34] M. Alice, "Fitting a Neural Network in R; neuralnet package," *datascienceplus*, 2018.

[35] K. Kim and J. Hong, "A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis," *Pattern Recognit. Lett.*, pp. 1–8, 2017, doi: 10.1016/j.patrec.2017.08.011.