Risk Prediction Model for Dengue Transmission using Artificial Neural Networks

Leslie Chandrakantha
Department of Mathematics & Computer Science
John Jay College of Criminal Justice of CUNY, USA

ABSTRACT
Dengue fever is a mosquito-borne viral disease that has grown dramatically around the world in recent years. It is more prevalent in tropical and subtropical countries. Annually, an estimated 390 million infections occur worldwide. Several studies have shown that climate factors influence this disease. Furthermore, it was shown that the influence of climate factors on dengue incidences was expected to be visible after some lag period. Identifying the climate factors that influence the spread of dengue fever would be helpful in combatting growth of the disease. This study builds an Artificial Neural Network (ANN) model for predicting the risk status of dengue incidences based on climate factors. The climate factors, average temperature, rainfall, and average relative humidity with a time lag are used as input parameters to the ANN. The monthly dengue incidences and the data on climate factors from the city of Colombo in Sri Lanka are used for this study. The accuracy of the ANN model prediction is found to be 90%.

General Terms
Artificial Neural Networks, Risk Prediction

Keywords
Dengue incidences, Artificial Neural Networks, Risk prediction, Climate factors

1. INTRODUCTION
Dengue fever is a mosquito-borne viral disease mostly prevalent in tropical and subtropical countries. It has become a major global burden for many countries. It is transmitted by the female mosquitoes of the *Aedes* type, mainly *Aedes aegypit* [1]. According to the Center for Disease Control and Prevention (CDC) of the United Sates, as many as 390 million people worldwide are infected annually and approximately 40,000 die [2]. About 2.5 billion people live in dengue affected countries and that include the tropical and subtropical areas in Africa, the Americas, and the Asia Pacific region [3]. In 2019, a significant increase in number of cases was seen. The direct and indirect expenses for dengue management and services are substantial and impose enormous burdens on low-and middle-income tropical countries, with a global estimate of US $8.9 billion per year [4]. Typically, the symptoms of dengue fever are similar to those of the flu. But the infection can produce a wide spectrum of illnesses which range from hemorrhagic manifestations, plasma leakage, and severe organ impairment with potentially lethal complications. Since there are no specific antiviral treatments or vaccines preventing dengue fever, the most effective way to manage the disease is through preventive medicine, control of mosquito population and avoiding mosquito bites [5].

This paper uses the data collected from Sri Lanka. Sri Lanka is one of the leading countries affected by the dengue fever outbreaks in recent years. Dengue infections have been a major health problem for many years in Sri Lanka. The first case of dengue fever serologically confirmed in Sri Lanka in 1962 [6]. The prevalence of dengue infections has been increasing on yearly basis overtime. According to the records' of Epidemiology Unit of Ministry of Health in Sri Lanka [7], on average, about 69,000 countrywide infections occurred during last eight years. About 23% of the total dengue incidences in the country are reported from Colombo district which is the leading district among 25 districts of the country. Colombo district is subdivided into two parts, namely, Colombo city (Colombo Municipal Council area) which is the capital of the country, and the surrounding cities and suburban areas. Data shows that nearly 25% of the dengue incidences in Colombo district occurred in Colombo city during last eight years. Based on these statistics, the population size and the importance of the location of Colombo city in the country, it would be appropriate to use the data from Colombo city for this paper.

The objective of this paper is to build a risk prediction model for dengue incidences based on climate factors. The artificial neural network (ANN) model is utilized for this purpose. The model predicts the likelihood of having high dengue incidences based on climate factors. A number of previous studies have demonstrated the relationship between dengue incidences and the climate factors. The majority of these studies used either time series or regression methods for modeling dengue forecasting. Morin et al. [8] noted that climate influences dengue ecology by affecting vector dynamics, agent development, and mosquito-human interactions. They further mentioned that relationships between climate variables and factors that influence dengue transmission are complex and future research will enable better projections of climate change effects on dengue incidence. Chandrakantha [9] identified that the rainfall data within two-month lag period were a significant predictor of dengue incidences in Colombo, Sri Lanka. Their work was based on Poisson and negative binomial regression modeling. Vu et al. [10] identified that temperature, relative humidity, sunshine, and rainfall had significant association with dengue incidences. Withanage et al. [11] used three time series forecasting models for dengue incidences with climate variables as predictors. They analyzed various lag times and noted that the previous month's dengue cases had a significant effect on the dengue incidences of the current month.

In the last decade, ANN models have been used in many disciplines, such as engineering, geography, and epidemiology to establish meaningful findings from the available data. Their ability to learn from given data makes these methods perfect tools since there is no need to make specific parametric assumptions or mathematical models. This gives an advantage over traditional statistical approaches [12]. ANN models have been successfully used for credit risk prediction in bank loans based on relevant factors [13,14]. Studies in Thailand, Singapore, and Malaysia have used ANNs to predict dengue fever cases with accuracies greater than 80% [15, 16, 17]. A

similar study in Sri Lanka using ANN models showed a lower accuracy [18]. Ughelli et al. [19] used ANN models to predict the number of dengue cases in Paraguay based on the relationships with climate variables. Their work concluded that different climate variables affect the number of cases for different districts.

This paper develops a risk prediction model for dengue incidences based on climate variables using artificial neural networks (ANN). In this case, the dependent variable of the model (risk status) will be a binary variable which indicates whether the dengue count is high (high risk) or low (low risk) for a given month. ANN models are frequently used in predicting outcomes of binary variables [20]. Monthly data were used for model training and testing. Using this approach, it is possible to identify if a specific month will be at risk for high dengue incidences based on the month's climate. This finding will be useful for authorities to create a dengue warning system.

This paper is organized as follows: Section 2 gives a brief introduction of artificial neural networks. Section 3 provides the methodology. Section 4 gives the results and a model evaluation. The conclusions are given in section 5.

2. ARTIFICILA NEURAL NETWORK

Artificial neural networks (ANNs) have been used as a modeling tool in all areas for decision making. They are commonly used as a standard nonlinear alternative to traditional linear approaches, especially in situations where the data possesses a complex relationship among variables [21]. ANN is a learning system that gathers their knowledge by detecting the patterns and relationships in data and learns (or is trained) through experience. They are also known as black box systems, in which extraction of information from internal system is impossible. The network is composed of a set of connected nodes called artificial neurons which are similar to biological neurons. Neurons are connected by links with associated weights which represents relative importance of the connection. The network has three types of layers, namely, the input layer, the hidden layer (middle layer), and the output layer. The number of hidden layers is usually one or two. The *Figure 1* shows the structure of an artificial neural network.

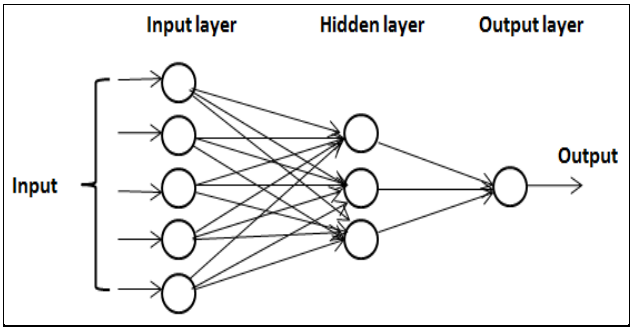

Figure 1: Basic Structure of ANN

The input layer has the input neurons which receives input stimulus. Then the input information is transferred to the next layer known as the hidden layer. The neurons between these two layers are connected with respective weights which represent the relative importance of the connection. This means that the information obtained from each neuron is sized according to the weight of the connection between the two neurons. The neurons within the same layer are not connected. The job of the hidden layer is to transform the information from input layer into something meaningful that the output layer can use in some way. A typical architecture of a neuron is shown in *Figure 2*.

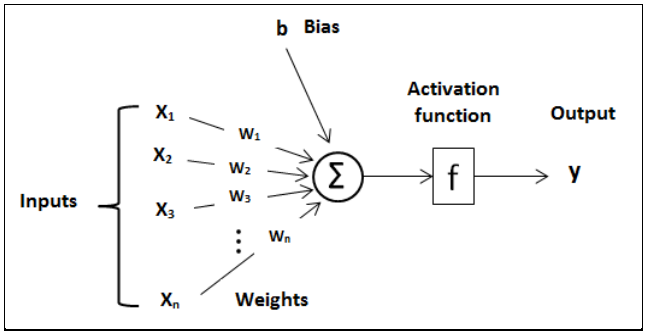

Figure 2: A Single Neuron

A neuron can have several inputs, but has only one output. Each neuron has the threshold value, the transfer function and the associated weights. The threshold is the minimum value that the input must have to activate the neuron. Hidden layer are the neurons that constitute the middle layer. For each neuron, it computes the summation of weighted sum and the bias, subtracts its threshold from this sum, and applies the transfer function (activation function). The output of the neuron is the result obtained from the transfer function. The bias is used to shift the transfer function up or down. The output of a neuron is a mathematical function of its inputs. The common transfer functions used in neural networks are the sigmoid, tanh, and Relu functions [22].

A major function of a neural network is the learning from exiting data. In this task, specific turning of the weights has to be done. It is accomplished by a learning algorithm which trains the network and modifies weights iteratively until desired output is computed. Typically, the learning algorithm stops when the error between the actual output and the desired output falls below a predefined threshold value.

There are three types of learning algorithms of artificial neural networks: supervised learning, unsupervised learning, and reinforced learning. In supervised learning, a training set of input-output pairs is used to train the network. The training set consists of pairs of inputs and desired outputs. A supervised learning algorithm analyzes the training data set and produces an inferred function which can be used in new input data. In supervised learning, a network produces outputs based on previous experience. Common applications of supervised learning are known as classification problems. In a classification situation, a network learns from the given data and makes new observations. In this paper we use the supervised leaning and classification to predict the risk status of dengue transmission. In unsupervised learning, only the input data is available. The algorithms guide the networks to find the important structure in the data. In reinforced learning, the network makes a sequence of actions by awarding rewards or penalties for each action. The goal is to maximize the total reward.

3. METHODOLOGY

3.1 Data Source

Data for this study was obtained from two Sri Lankan sources. The monthly dengue counts in the city of Colombo from 2010 to 2019 were obtained from the epidemiology department of Ministry of Health of Sri Lanka [8]. The monthly climate data in the city of Colombo for the same time period were extracted from the yearly statistical abstracts of the Department of Census and Statistics [23]. This climate data includes monthly average temperature (^{o}C), cumulative rainfall per month (mm), and monthly average relative humidity. The total number of data points used in this study is 120.

This paper aims to predict whether a given month would be at risk for high dengue incidences. The Artificial Neural Network approach is used for the model formulation. The dependent variable, the risk status, is defined as whether the monthly dengue incidences are above the median dengue incidences (1) or not (0) during the period of 2010 to 2019. In this context, a month will be at for high risk if the dengue incidences are above the median dengue incidences for that period or otherwise the month will not be at high risk. The neural network differentiates a month between riskier for high dengue incidences or not by using 0 and 1. The independent variables are the average temperature, cumulative rainfall and average relative humidity. *Figure 3* shows the number of dengue incidences for each month from 2010 to 2019. It clearly shows several outliers. The median was chosen as the threshold of risk since it was not influenced by outliers.

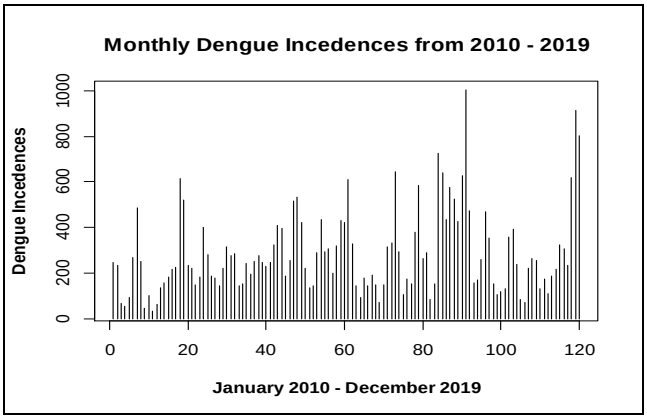

Figure 3: Monthly Dengue Incidences

The paper models the relationship between risk status and the climate variables using two months lagged climate data. The reason for this is that it takes a certain time period for an egg to develop into an adult mosquito and the influence of climate is expected to be visible after one or two months [24]. For one lag-month data, the Pearson correlation coefficients between dengue incidences and climate data were 0.1641 ($p > 0.05$), -0.0764 ($p > 0.05$) and 0.2380 ($p < 0.05$) respectively for rainfall, average temperature and average relative humidity. For two lag-months data, the correlation coefficients were 0.4352 ($p < 0.05$), 0.0339 ($p > .05$) and 0.3419 ($p < 0.05$), indicating significant positive correlation between dengue incidences and two of the three climate variables.

3.2 Model

The *neuralnet* function in R [25] is used to fit the ANN model for predicting the risk status of dengue incidences. This fitted network has an input layer with three input variables (three climate variables noted earlier), two hidden layers and one output layer which compute the risk status. The network is trained by using a supervised learning algorithm (back propagation algorithm). The algorithm optimizes the neuron weights by minimizing the error between the actual and desired output. The algorithm will work until a stopping criterion is found [26].

The entire dataset is divided in to two portions, the training set and the test set. The training set which consists of 75% of the data, used to train the ANN, while the test set is used to validate the performance of the model. The data normalization is performed before inputting data into the network to ensure that the data range is in the same interval. This will allow the network to learn optimal parameters more quickly for each input node [27]. The max-min linear transformation function was used to normalize the data before splitting the dataset into training and test datasets. As you can see from *Figure 4*, there are three input nodes, two hidden layers and one output node. The training process took 821 steps for the convergence with an error of 8.79.

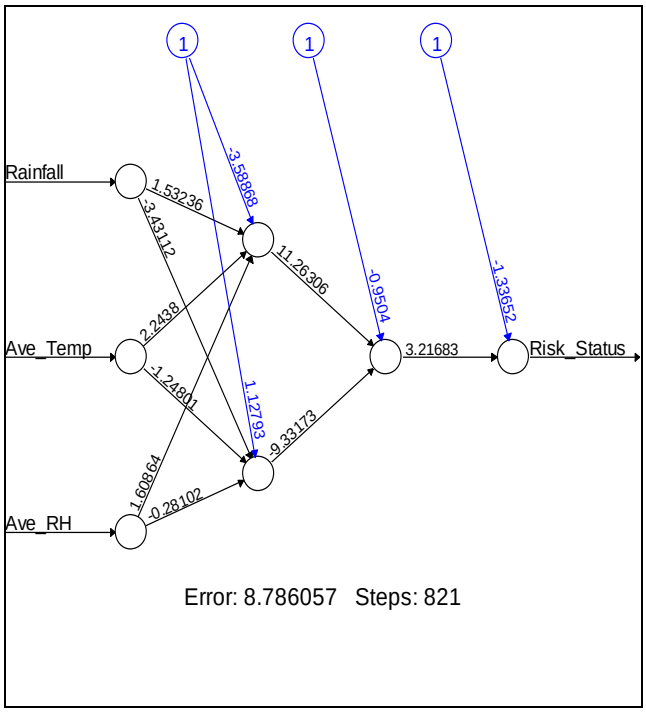

Figure 4: Fitted Neural Network with Three Input Variables

4. RESULTS AND MODEL EVALUATION

First, the training data set was used to train the network and then, it was tested using the test data set. The *compute* function in R is used to compute the output for new inputs. The output of the network is a classifier that results 0 and 1. *Figure 5* shows a portion of the output (predicted risk status) produced by the ANN model with actual values given in test data set. The entire output resembles the same pattern shown in the portion of the output in *Figure 5*. The most of the predicted values agree with the actual values of the risk status. This indicates that our ANN model is predicting well.

actual	prediction
1	1
1	1
0	0
0	0
1	1
1	1
1	0
0	0
0	0

Figure 5: Portion of the Output with Actual Values

The graphical visualization of generalized weights is a way to examine the relative contribution of each input variable. The generalized weights are used to study the effects of individual input variables for predicting the output. A large variance in generalized weight of an input variable indicates a nonlinear effect on the output variable. If the generalized weight of a covariate is approximately zero, the covariate has little effect on the output [28]. *Figure 6* shows the generalized weights for the

three input variables. Larger variances in generalized weights for all three inputs show a nonlinear effect on risk status.

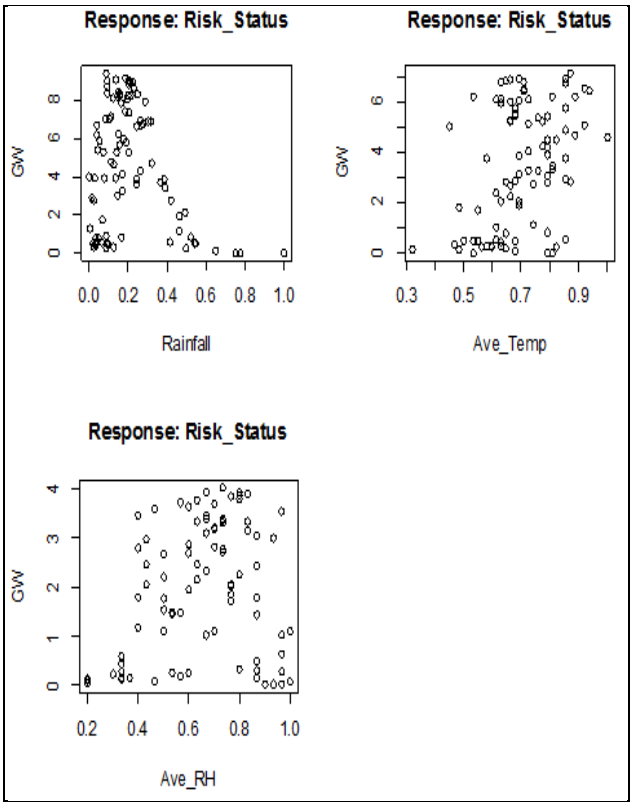

Figure 6: Generalized Weights for the Input

A confusion matrix is a table that can be used to measure the performance of the algorithm used in ANN. This table tells us an idea about how well the classifier in the ANN model has performed. The table is normally computed using the test set and it lists the actual values and predicted values for the outcome. *Figure 7* shows the confusion matrix for this test data set. The accuracy of the prediction is calculated as the number of all correct predictions divided by the total number of data in test set. Based on this confusion matrix, the accuracy rate is 90% and the error rate is just 10%. This accuracy rate provides the evidence that the ANN model has performed well.

	prediction	
actual	0	1
0	14	2
1	1	13

Figure 7: Confusion Matrix

Figure 8 shows a sketch of the ROC (Receiver Operating Characteristic) curve for this network. It is one of the most important and popular evaluation metrics for checking any classification model's performance. Higher the area under the curve (AUC), better the model is at predicting 0s as 0s and 1s as 1 [29]. The area under the curve was 0.9018, with 95% confidence interval ranging from 0.7927 to 1.0. This is in the range of good to excellent prediction performance.

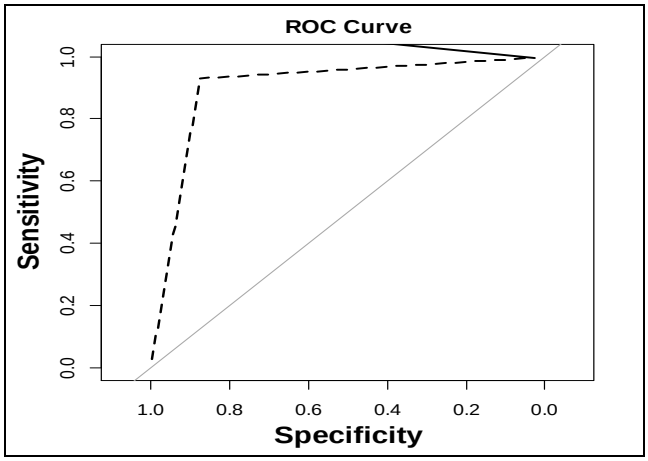

Figure 8: Receiver Operating Characteristic (ROC) Curve

Based on the above measures the ANN has performed well in predicting the risk status of dengue occurrence for a given month based on climate variables. The generalized weights have shown that all three climate factors affect the dengue occurrence. The confusion matrix and the ROC curve were used to evaluate the accuracy of the prediction of the outcome. Those accuracy rates are at 90%.

5. CONCLUSIONS

Dengue fever has been a critical public health problem for many countries. This paper used the artificial neural network to predict the risk status (high or low) based on climate factors. The neural network has been trained on data from city of Colombo. The data set contained monthly data from 2010 to 2019. The results have shown that the accuracy of prediction is 90%. The advantage of use of ANN models is that there is no need to make any specific parametric assumptions or mathematical models. Their ability to learn from given data makes ANN models perfect tools for predicting future outcomes. In summary, the fitted neural network provides strong evidence to efficiently predict the risk status of dengue incidences based on climate data. This proposed prediction model can be used worldwide. These findings are helpful for authorities to take necessary actions in safeguarding the community from dengue outbreaks. This model can be expanded using additional variables that relate to mosquito growth.

6. REFERENCES

[1] WHO (World Health Organization): Who spreads dengue and severe dengue? https://www.who.int/denguecontrol/faq/en/index5.html, accessed 05 August 2010.

[2] CDC. Centers for disease control and prevention, https://www.cdc.gov/dengue/index.html, accessed 01 August 2020.

[3] WHO (World Health Organization) 2009: WHO Report on Global Surveillance of Epidemic-prone Infectious Diseases - Dengue and dengue hemorrhagic fever. https://www.who.int/csr/resources/publications/dengue/CSR_ISR_2000_1/en/index5.html, accessed 01 August 2020.

[4] Shepard, D. S., Undurraga, E.A., Hallasa, Y. A., and Stanaway, J. D. 2016. The global economic burden of dengue: a systematic analysis. *Lancet Infectious Diseases*, 16, 935–941

[5] Al-Muhandis, N. and Hunter, P. R. 2011. The Value of

Educational Messages Embedded in a Community-Based Approach to Combat Dengue Fever: A Systematic Review and Meta Regression Analysis. *PLOS Neglected Tropical Diseases,* *5*(8): e1278. https://doi.org/10.1371/journal.pntd.0001278

[6] Sirisena, P.P.N.N. and Noordeen, F. 2014. Evolution of Dengue in Sri Lanka—Changes in the Virus, Vector, and Climate. *Int. J. Infect. Dis*., 19, 6–12

[7] Epidemiology Unit of Ministry of Health of Sri Lanka. Available Online: http://www.epid.gov.lk/web/

[8] Morin, C. W., Comrie, A. C. and Ernst, K. C. 2013. Climate and dengue transmission: evidence and implications. *Environ Health Perspectives*. 121, 1264–1272; http://dx.doi.org/10.1289/ehp.1306556

[9] Chandrakantha, L. 2019. Statistical analysis of climate factors influencing dengue incidences in Colombo, Sri Lanka: Poisson and negative binomial regression approach. *Int. J. Sci. Res. Publ.* 9, 133–144, doi:10.2322/IJSRP.9.02.2019.p8616

[10] Vu, H.H., Okumura, J., Hashizume, M., Tran, D.N. and Yamamoto, T. 2014. Regional differences in the growing incidences of dengue fever in Vietnam explained by weather variability. *Trop. Med. Health*. 42, 25–33. doi: 10.2149/tmh.2013-24

[11] Withanage, G.P., Wishwakula, S.D., Gunawardena, Y.I and Hapugoda, M.D. 2018. A Forecasting Model for Dengue Incidence in the District of Gampaha, Sri Lanka. *Parasites and Vectors*. 11, 262, doi:10.1186/s13071-018-2828-2.

[12] Breiman, L. 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*. 16(3): 199–231

[13] Pacelli, V. and Azzollinni, M. 2011. An Artificial Neural Network Approach for Credit Risk Management, *Journal of Intelligent Learning Systems and Applications*, 3, 103-112

[14] Gupta, D. P. and Goyal, S. 2018. Credit Risk Prediction Using Artificial Neural Network Algorithm, *I.J. Modern Education and Computer Science*, 5, 9-16

[15] Rachata, N., Charoenkwan. P., Yooyativong. T., Chamnongthai . K., Lursinsap, C. and Higuchi, K. Automatic prediction system of dengue haemorrhagic-fever outbreak risk by using entropy and artificial neural network; *Proceedings of the International Symposium on Communications and Information Technologies*; Vientiane, Laos. 21–23 October 2008

[16] Aburas, H. M., Cetiner , B. G. and Sari, M. 2010. Dengue confirmed-cases prediction: A neural network model. *Expert Syst. Appl.* 37, 4256–4260. doi: 10.1016/j.eswa.2009.11.077

[17] Hwang, S., Clarite, D.S., Elijorde, F.I., Gerardo, B.D. and Byun, Y. 2016. A web-based analysis of dengue tracking and prediction using artificial neural network. *Advanced Science and Technology Letters*; Science & Engineering Research Support Society; Sandy Bay, TAS, Australia: 122, 160–164.

[18] Nishanthi, P., Perera, A. and Wijekoon, H. 2014. Prediction of dengue outbreaks in Sri Lanka using artificial neural networks. *Int. J. Comput. Appl*., 101, 1–5

[19] Ughelli, V., Lisnichuk, Y., Paciello, J., and Pane, J. 2017. Prediction of Dengue Cases in Paraguay Using Artificial Neural Networks. *Int'l Conf. Health Informatics and Medical Systems – HIMS 1*, 130-136

[20] Ong, E. and Flitman, A.1997. Using neural networks to predict binary outcomes, *IEEE International Conference on Intelligent Processing Systems* (Cat. No.97TH8335), Beijing, China, 1, 427-431 doi: 10.1109/ICIPS.1997.672816.

[21] Bertolaccini, L., Solli, P., Pardolesi, A., and Pasini, A. 2017. An overview of the use of artificial neural networks in lung cancer research. *Journal of thoracic disease*, 9(4), 924–931. https://doi.org/10.21037/jtd.2017.03.157

[22] Maca, P., Pech. P. and Pavlasek, J. 2014. Comparing the Selected Transfer Functions and Local Optimization Methods for Neural Network Flood Runoff Forecast, *Mathematical Problems in Engineering*. https://doi.org/10.1155/2014/782351

[23] Department of Census and Statistics of Sri Lanka. http://www.statistics.gov, accessed 01 August 2020.

[24] Nakhapakorn, K. and Tripathi, N.K. 2005. An Information Value Based Analysis of Physical and Climatic Factors Affecting Dengue Fever and Dengue Haemorrhagic Fever Incidence. *Int. J. Health Geogr. 4*, doi:10.1186/1476-072X-4-13.

[25] Gunther, F. and Fritsch, S. 2010. neuralnet: Training of Neural Networks, *The R Journal.* 2(1), 30-38. https://journal.r-project.org/archive/2010/RJ-2010-006/RJ-2010-006.pdf

[26] Pineda, F. J. 1987. Generalization of back-propagation to recurrent neural networks. *Physical review letters*, 59(19), 2229.

[27] Nayak, S. C., Misra, B. B. and Behera, H. S. 2014. Impact of Data Normalization on Stock Index Forecasting, *International Journal of Computer Information Systems and Industrial Management Applications*. 6, 257-269.

[28] Zhang, Z. 2016. Neural networks: further insight into error functions, generalized weights and others. *Annals of Translational Medicine*. 4(16), 300. doi: 10.21037/atm.2016.05.37.

[29] Woods, K. and Bowyer, K. W. 1997. Generating ROC curves for artificial neural networks, *IEEE Transactions on Medical Imaging*, 16 (3), 329-337. doi: 10.1109/42.585767.