Risk Prediction of Dengue Transmission using Artificial Neural Networks and Regression

Leslie Chandrakantha
Department of Mathematics and Computer Science
John Jay College of Criminal Justice of CUNY, USA
lchandra@jjay.cuny.edu

Abstract
Dengue fever is fast emerging mosquito-borne viral disease with more than one third of the world's population is at stake. The objective of this paper is to predict the risk status (high or low) of dengue transmission based on climate factors using artificial neural networks and the logistic regression method. Previous studies have shown that climate factors influence dengue transmission. The rainfall, temperature, and relative humidity are considered as input for both models. The artificial neural networks possess the ability to identify complex relationships in data without any specific assumptions. Both models fared well and showed minimal false predictions.

1. Introduction

Dengue fever is spread by mosquitos. Understanding the nature of dengue fever and identifying the preventative measures is essential for addressing the spread and outbreaks of the disease. It is mostly prevalent in tropical and subtropical countries. According to the Center of Disease Control and Prevention (CDC) of the Unites States, as many as 390 million people worldwide are infected annually [1]. Most of these infections occur in tropical and subtropical areas in Africa, the Americas, and the Asia Pacific region [2]. In recent years, a significant increase in the number of cases was seen. The symptoms of dengue fever are similar to those of the flu. The fatality rate is normally lower than 1%, but because of the absence of proper diagnosis and treatment, it can be as high as 20% [3]. Since there is no antiviral treatment for dengue fever, controlling the mosquito population and avoiding mosquito bites are key preventative measures.

In this paper, we use data collected from Colombo, the capital city of Sri Lanka. Sri Lanka is one of the countries severely affected from dengue outbreaks in the recent years. The objective of this paper is to study the ability of the artificial neural network (ANN) algorithms and logistic regression methods to predict the risk status (whether high or low) of dengue transmission based on climate factors. The two models predict the likelihood of having high dengue incidences and interpret it as the risk status. A number of previous studies have investigated the relationship between dengue incidences and climate factors. Many of them concluded that there was an association between dengue incidences and climate factors. Vu et al. [4] identified that levels of temperature, relative humidity, sunshine, and rainfall were correlated with dengue incidences. Withanage et al. [5] used time series forecasting models and concluded that the previous month's dengue cases and climate factors had a significant effect on current month's dengue incidences. Using Poisson and negative binomial regression modeling, Chandrakantha [6] identified that the

rainfall data within a two-month lag period was a significant predictor of dengue incidences in Colombo, Sri Lanka.

In recent years ANN models have been used to establish meaningful relationships in available data sets in many different applications. They are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found [7]. One significant advantage is the ability to learn from the data itself. Pacelli and Azzollini [8] have used ANN models in predicting credit risk for Italian companies. They have concluded that neural networks represent an alternative to traditional methods of classification because they are adaptable to complex situations. Ughelli et al. [9] 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.

In this paper, we develop risk prediction models for dengue incidences based on climate variables using an ANN model and a logistic regression model. The dependent variable of the models, the 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 and logistic regression models are frequently used in predicting outcomes of binary variables [10, 11]. In the ANN model, monthly data was used for model training and testing. Using these two approaches, it is possible to identify whether a specific month will be at risk for high dengue incidences based on the month's climate. The accuracies of both models will be compared. These findings 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 and logistic regression. Section 3 provides methodology. Section 4 gives the results and model evaluations. We end the paper in Section 5 with conclusions.

2. Artificial Neural Networks and Logistic Regression

2.1 Artificial Neural Networks

Artificial neural networks (ANN) mimic the human brain in processing input signals and transforming them into output signals. Their algorithms allow nonlinearity between input variables and output variables [12]. An ANN model is suitable for a situation where the investigator does not know the underline functions. In other words, an ANN is able to learn from data without any specific functional assumptions. This learning process is known as training of the network. The network is composed of a set of connected nodes called artificial neurons. Artificial neurons 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. *Figure 1* shows the structure of an ANN.

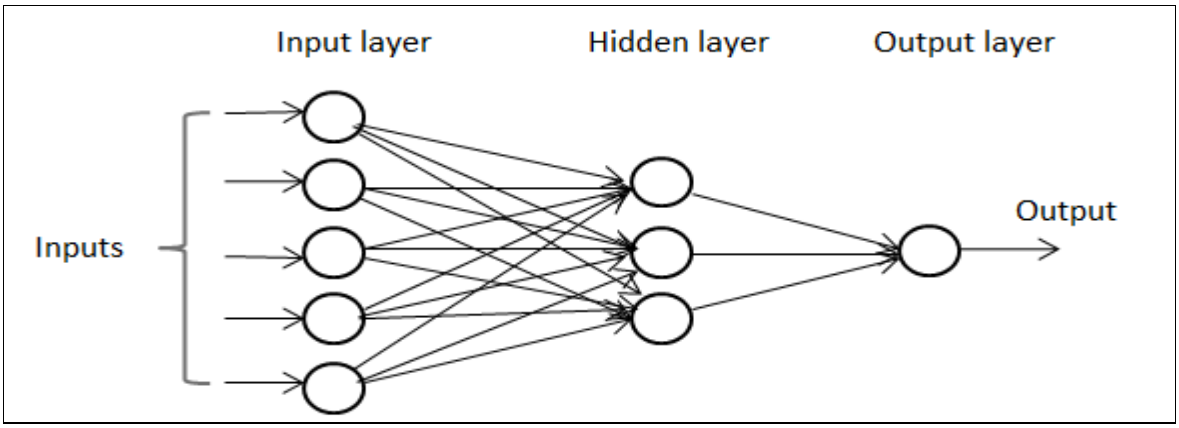

Figure 1: Basic Structure of ANN

The input layer has the input neurons which receives input stimuli. 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 the 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*.

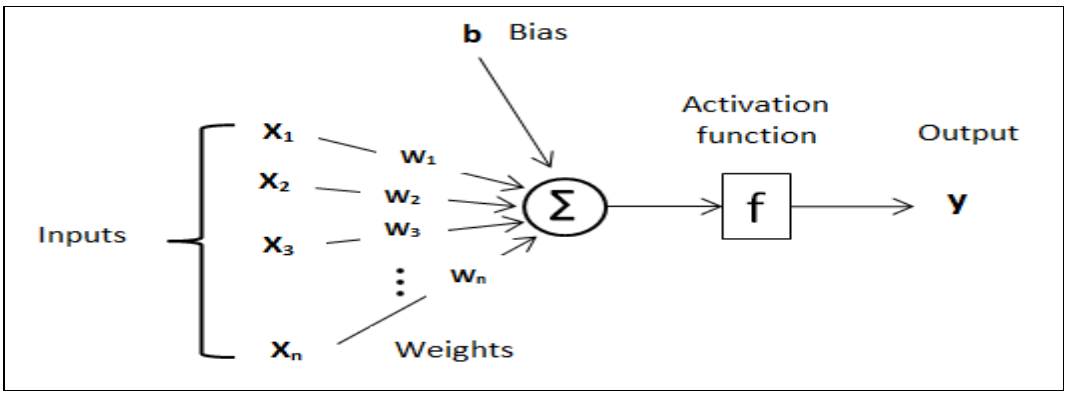

Figure 2: A single Neuron

A neuron can have several inputs, but has only one output. Each neuron has a threshold value, a transfer function and associated weights. The threshold is the minimum value that the input must have to activate the neuron. The hidden layer is the neurons that constitute the middle layer. *All of the input variables are combined across one or more nodes in the hidden layer.* This essentially creates new features that are derived from the input. The hidden layer performs computations on the weighted inputs and produce net input which is then applied with the transfer function (activation function) to produce the actual output. For each neuron, it computes the summation of weighted sum and the bias, subtracts its threshold from this sum, and applies the transfer 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 output computation procedure can be mathematically expressed as

$$y(x) = \emptyset\left(\sum_{i=1}^{n} w_i x_i\right)$$

where y is the output signal, $\phi(\ )$ is the activation function, x is input variables and w is weight assigned to each input variables. The common transfer functions used in neural networks are sigmoid and tanh functions.

A major function of a neural network is learning from existing 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 a 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. In this study, we are using a supervised learning algorithm. In supervised learning, a training set of input-output pairs is used trained 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.

2.2 Logistic Regression

The logistic regression model is a widely used binary data modeling approach that belongs to the family of generalized linear models (GLM) [13]. The logistic regression is used to model the relationship between a binary response variable and a set of predictors that can be discrete, continuous, or categorical. Since the response variable is binary, it can take a value of either 0 or 1.

For a binary response variable Y with, $Y = 0$ or 1, and a single predictor variable x, we write $E(Y|x) = P(Y = 1\ |x)$ as a function of x as follows: First we abbreviate $p(x) = P(Y = 1|x)$ and then write $p(x)$ as

$$P(Y = 1|x) = p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

This is known as the logistic regression model. Rearranging this equation gives:

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

The left hand side of the expression above is called the log odds or logit of $p(x)$. The expression $\frac{p(x)}{1-p(x)}$ gives the odds of the event $Y = 1$ occurring. For multiple predictor variables $X = x_1, x_2, x_3, \ldots, x_k$, the logistic regression model becomes:

$$P(Y = 1|X) = p(X) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

and the logit becomes

$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

From the above expression, we notice that β_i is the change in log odds of $Y = 1$ occurring for one unit increase in x_i while holding other predictors are fixed. The simple algebra can be used to show that e^{β_i} is the odd ratio associated with a one unit increase in predictor x_i. An interpretation of this odds ratio is that for each 1 unit increment of x_i, holding other predictors are fixed, the percentage change (increase/decrease) of odds for Y occurring is e^{β_i}-1.

3. Methodology

Data for this study was obtained from 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 [14]. 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 [15]. This climate data includes monthly average temperature (°C), cumulative rainfall per month (mm), and monthly average relative humidity.

In this paper, we predict whether a given month would be at risk for high dengue incidences. The dependent variable is the risk status. The risk status is defined as whether the monthly dengue incidences were above the median dengue incidences (1) or not (0) during the period of 2010 to 2019. In this context, a month is at a high risk if the number of dengue incidences is above the median dengue incidences for that period. If not, the month is not at high risk. The independent variables are the monthly cumulative rainfall, average temperature, and average relative humidity. The median is chosen as the threshold for risk since it is not influenced by outliers. Since the dependent variable assumes either 0 or 1, both the ANN model with binary classification and the logistic regression model can be used for risk prediction.

Since the influence of climate factors on dengue incidences is visible after some time period, we modeled the relationship between the risk status and the climate factors using lagged data from two months ago. The correlation between dengue incidences and two months lagged climate variables show a significant positive correlation while one month lagged data and current month data do not show a significant positive correlation.

3.1 ANN Model

The fitted network has an input layer with three input variables corresponding to climate variables, two hidden layers and one output layer that gives risk status likelihood. 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 [16]. The entire data set is divided into two parts, the training set (75% of the data) and the test set (25% of the data). The training set is used to train the network and 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 [17]. The max-min linear transformation function was used to normalize the data before splitting the dataset into training and test datasets.

We used the *neuralnet* function in R [18] to train the ANN model. The R statement for implementing the *neuralnet* function is given below:

```
> install.packages("neuralnet")
> library(neuralnet)
> nn <- neuralnet(Risk_Status ~ Rainfall + Ave_Temp + Ave_RH, data = mydata, hidden =
c(2,1), linear.output = FALSE)
> plot(nn)
```

The first two lines of the code shown above install the *neuralnet* package and then load it to the working space. The first argument of the *neuralnet* function describes the model to be fitted. Risk_Status is the response (output) variable. The predictors are connected by "+" symbols. The *data* argument specifies the data frame that contains the data for training. The parameter *hidden* is a vector which specifies the number of hidden layers and the number of units in each layer. The *linear.output* parameter is set to FALSE to indicate that this is a classification situation. The default activation function which is the logistic function is used in this case. The topology of network model can be visualized using the *plot* function. *Figure 3* displays the topology of the ANN model. It shows that the training process took 331 steps for the convergence with an error of 11.26.

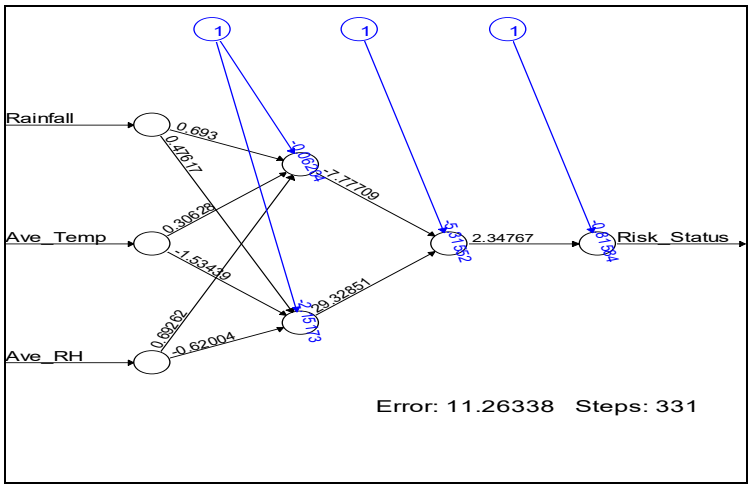

Figure 3: Fitted ANN Model

For predictions, only input data is needed. For this purpose, we remove the risk status data column from the test set. Now it only contains the climate data of the test set. Let's suppose this data is stored in data frame named *tempest*. The probability having high dengue counts (high risk) can be calculated using the *compute* function as shown below:

> nn.results <- compute(nn,tempest)

Now the data frame *nn.results* contains the probabilities. These probabilities are rounded to 0 or 1 to obtain the risk status. This way we can predict the risk status for dengue incidences for any given month with the values from three climate variables.

3.2 Logistic Regression Model

The logistic regression model can be created using *glm* function in R. The flowing is the R statement to create the model:

> mylogit = glm(Risk_Status ~ Rainfall + Ave_Temp + Ave_RH, data = mydata, family = "binomial")

The first argument is the same as in the ANN model creation. In this case we use original data without normalizing. The *family* parameter is specified as "binomial" to use logistic regression. To compute the predicted probabilities, the *predict* function can be used as follows:

> mylogit.results <- predict(mylogit, newdata = tempset, type = "response")

These probabilities are also rounded to 0 or 1 to obtain the risk status for the logistic regression model.

4. Results and Model Evaluations

The test data set is used to compute the risk status values for both models. *Figure 4* shows a portion of the output with actual values for both models.

Actual	ANN	Logit Regression
1	1	1
1	1	1
0	0	0
0	0	0
1	1	1
1	1	1
1	0	1
0	0	0

Figure 4: Portion of the Output with Actual Values

In ANN model, the generalized weights are used to study the effects of individual covariates for predicting the output. A large variance in the generalized weight of a covariate 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 [19]. *Figure 5* shows the generalized weights for three input variables. Larger variances in generalized weights for all three inputs show a nonlinear effect on risk status.

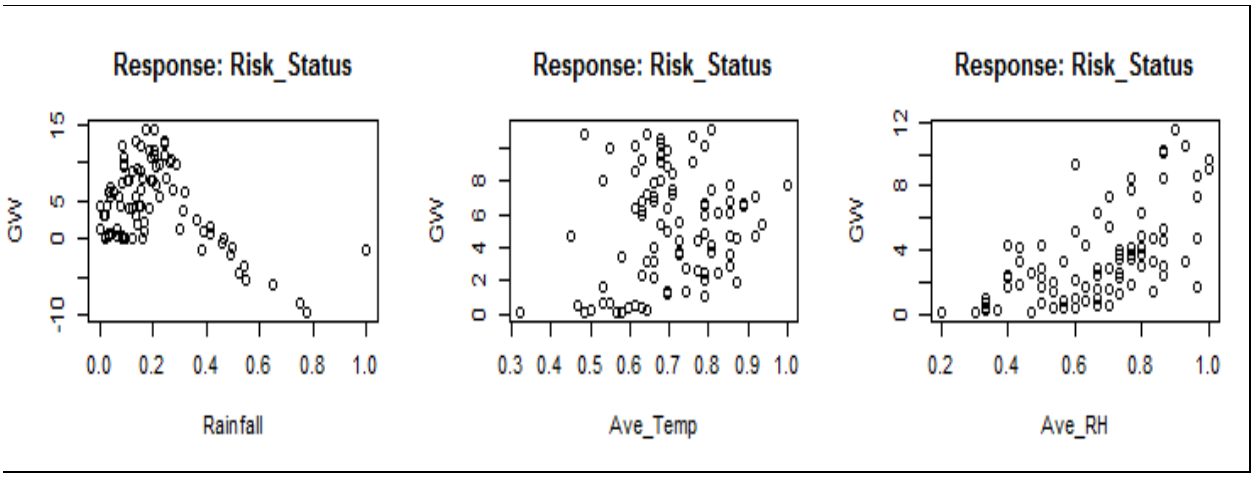

Figure 5: Generalized Weights

We computed the Cook's distances for the logistic regression model to identify any influential observations. Cook's distance measures the overall influence an outlying observation has on the estimated coefficients of the regression model [20]. The Cook's distances can be compared with values of the F distribution with $(k+1)$ and n - $(k+1)$ degrees of freedom, respectively, to identify the influence. Usually, an observation with a value of the Cook's distance that falls above the 50th

percentile of the *F* distribution is considered to be an influential observation. The largest Cook's distance for this dataset is 0.618 and the 50^{th} percentile of the *F* distribution with 4 and 116 degrees of freedom is 0.844. This indicates that there are no influential observations in this dataset.

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data. In this case the true values are known. This table lists the actual values and predicted values for the outcome. Since the output in the logistic regression model is also 0s and 1s, we generated the confusion matrix for this model also. *Figure 7* shows the confusion matrix for this test data set for both models. The accuracy of the prediction is calculated as the number of all correct predictions divided by the total number of data in the test set. Based on these confusion matrices, the accuracy rates are 90% and 93% respectively for the ANN model and the logistic regression model.

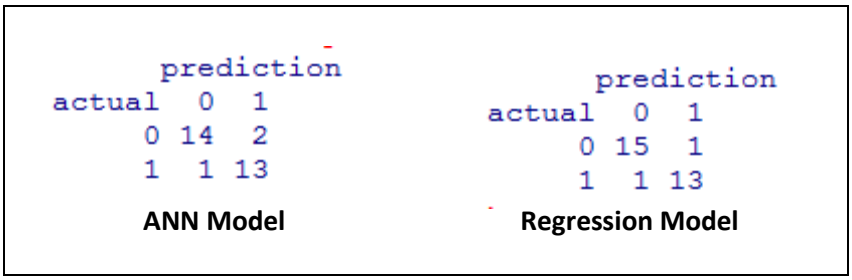

Figure 6: Confusion Matrix for Both Models

Figure 7 shows sketches of the ROC (Receiver Operating Characteristic) curves for both models. The ROC curve is an important and popular model evaluation metric for checking any classification model's performance. The higher the area under the curve (AUC), the better the model is at predicting 0s as 0s and 1s as 1s. The area under the curve for the ANN model was 0.9018, with a 95% confidence interval ranging from 0.7927 to 1.0. The same measure for the logistic regression model was 0.9330 with the confidence interval ranging from 0.84 to 1. Both measures are in the range of excellent prediction performance.

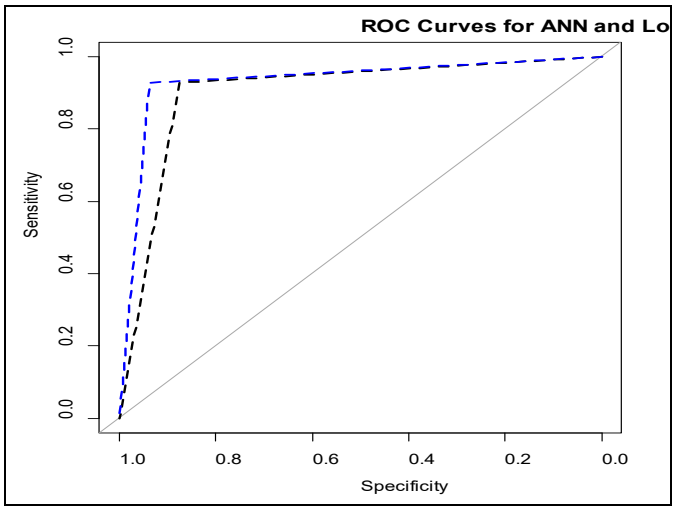

Figure 7: ROC Curves

Based on the above prediction performances, both ANN and logistic regression models performed well in predicting risk status of dengue incidences for a given month. The logistic regression model performs slightly better than the ANN model. In general, ANN models can be thought of as generalization of logistic regression models and they perform better than logistic regression models [21]. In some cases, ANN models can perform worse than logistic regression models because neural networks are more difficult to train and are more prone to overfitting than logistic regression.

5. Conclusions

Many countries have been severely affected by dengue fever outbreaks in the recent years. In this paper we used two methods, namely the artificial neural network model and the logistic regression model to predict the risk status based on climate factors. The neural network has been trained on data from the city of Colombo. The data set contained monthly data of dengue incidences and climate factors in the city of Colombo from 2010 to 2019. The results have shown that the accuracy of prediction for both models exceeds 90%. The advantage of using 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. Both models provide strong evidence to efficiently predict the risk status of dengue incidences based on climate data. These proposed prediction models can be used worldwide. These findings are helpful for authorities so that they can take necessary actions in safeguarding the community from dengue outbreaks. In this study, we have only considered three risk factors that affect dengue transmission: namely temperature, rainfall, and relative humidity. Some other factors can also affect the dengue transmission. We plan to include more risk factors in a future study.

References

[1] CDC. (Centers for disease control and prevention). https://www.cdc.gov/dengue/index.html

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

[3] Gubler, D.J. (1998). Dengue and dengue hemorrhagic fever. *Clinical microbiology reviews*. 11(3), 480-496. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC88892/

[4] Vu, H.H.,Okumura, J., Hashizume, M., Tran, D.N. & 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.

[5] Withanage, G.P., Wishwakula, S.D., Gunawardena, Y.I & 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.

[6] 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

[7] Mahmoodi, K. & Ketabdari, M.. (2016). Modeling the Slump Test and Compressive Strength of High-Performance Concrete Using Artificial Neural Networks and Multiple Linear Regressions. *Sharif Journal of Civil Engineering*. 33(2), 105-115

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

[9] Ughelli, V., Lisnichuk, Y., Paciello, J., & 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

[10] Ong, E. & 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.

[11] Dutta, A., Bandopadhyay, G. & Sengupta, S. (2015). Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression. *International Journal of Business and Information*. 7(1), 105-136

[12] Landi, A., Piaggi, P., Laurino, M. & Menicucci, D.. (2010). Artificial Neural Networks for nonlinear regression and classification. *ISDA 2010*. 115-120. 10.1109/ISDA.2010.5687280.

[13] Hosmer, D.W. Jr; & Lemeshow, (2013). S. Applied Logistic Regression, 3rd ed. John Wiley & Sons, Inc., New York.

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

[15] Department of Census and Statistics of Sri Lanka. http://www.statistics.gov

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

[17] Nayak, S. C., Misra, B. B. & 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.

[18] Gunther, F. & 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

[19] Intrator, O. & Intrator, N. (1999). Interpreting neural network results: A simulation study. *Computational Statistics and Data Analysis*. 37, 373-393.

[20] Mendenhall, W. & Sincich, T. (2003). A Second Course in Statistics: Regression Analysis, Sixth Edition, Pearson Education. Inc. New Jersey.

[21] Ayer, T., Chhatwal, J., Alagoz, O., Khan, C. Woods, R. & Burnside, E. (2010). Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. *RadioGraphics*. 30, 13-22.