Towards an Early Warning System to Combat Dengue

Aloka. Munasinghe, H.L. Premaratne, and M.G.N.A.S. Fernando

***Abstract*—**Prediction of dengue outbreaks is a timely need for Sri Lanka due to the dramatic increase of dengue incidences. Objective of this research is to create a simulation model to predict the future dengue outbreaks, intensity of the disease, and the population at risk of having the fever. The proposed model uses three main components, a regression model for identifying the risk factors that causes an outbreak, an Artificial Neural Network model for predictions, and a GIS model for simulating the results. Climatic, socio-economic, and census factors have been considered for predictions.

***Keywords*—**Artificial Neural Network, Dengue outbreak, GIS, Regression model.

I. Introduction

VECTOR-BORNE diseases have become a severe health problem in today's world and according to the World Health Organization (WHO) dengue is the most prevalent mosquito-borne viral disease among humans [1]. Every year 50 to 100 million dengue cases, 24000 deaths and 5 lacks of hospitalized cases are occurred globally and death rate is high especially among children. Disease is transmitted by the bites of infective female mosquitoes of the Aedes family, Aedes Aegypti or more rarely by Aedes Albopictus. Dengue is commonly found in South East Asia, Africa, USA, Western Pacific and Caribbean regions.

Dengue and the potentially fatal dengue hemorrhagic fever have been widely spread in Sri Lanka after the year 2000 and it has become a serious health problem to public health authorities in Sri Lanka. There has been a dramatic increase of dengue incidences in Sri Lanka in the past decade, mostly in urban and semi urban areas. During the year 2012 there have been 40144 dengue cases and 131 deaths in Sri Lanka [2]. However, currently there is no vaccine found for the disease. Developing a vaccine against the disease is said to be challenging as there are four closely related viruses causing the disease and the vaccine must be immune to all four serotypes.

Preventing occurrences of dengue outbreaks has become a significant problem in Sri Lanka. Early recognition of an outbreak supports the health officers to plan pre-emptive measures. However, current dengue outbreak prediction in Sri Lanka is incomplete and inefficient. The epidemiology unit uses three indices to monitor the dengue virus transmission namely House Index, Container Index and the Bretau Index. However, all these indexes are based upon the immature stage of the mosquitoes. Moreover, in the case of an outbreak the behavior of the disease is very uncertain.

There are several methods used to utilize in predicting a future outbreak in the proposed model. The research attempts to

Create a classification model to identify the risk factors

Identify the precise relationship between risk factors and outbreaks

Create a simulation model to predict outbreaks

The proposed model in the study considers the epidemiological, metrological, and census data for predictions.

Aloka. Munasinghe,is with University of Colombo School of Computing, Reid Avenue, Colombo 7, Sri Lanka (e-mail: aloka. ucsc@gmail.com).

H.L. Premaratne, is with University of Colombo School of Computing, Reid Avenue, Colombo 7, Sri Lanka (e-mail: hlp@ucsc.lk).

M.G.N.A.S. Fernando is with University of Colombo School of Computing, Reid Avenue, Colombo 7, Sri Lanka (e-mail:nas@ucsc.lk).

II. Related Work

Many researches that has been done before has proven that outbreaks can be predicted accurately using climatic and census factors. However, there are only a handful of studies that has been done based in Sri Lanka.

Most of the previous work done in this area is based on statistical or machine learning techniques. Regression analysis is a statistical method used to identify the relationship between one or more independent variable with a single dependent variable. Regression methods have been widely used in predicting and forecasting dengue outbreaks [4], [5], [6]. Pathirana *et al.* have done a study to assess the quantitative statistical relationship between dengue incidences and rainfall in Western province, Sri Lanka between 2000 and 2004. The correlation has been used to predict an outbreak. A regression model has been obtained using statistical analyses. The obtained association is in the form of

$$Ra\inf all = 13.588 \times \left(Cases \big/ Ra\inf all \right)^{-0.5018} \quad (1)$$

Number of disease outbreaks is expressed in terms of rainfall. The correlation coefficient for this power model has been obtained to be very high with a value of R^2=0.6362. It has shown that there is a three to four week lag time between the rainfall and outbreaks.

Geographically Weighted Regression (GWR) which is an ordinary least square regression method has been used to get the correlation between dengue incidences and population density in Johor State, Malaysia [6]. It states that the spatial distribution of the fever is highly correlated to the population density (R^2=0.87). Furthermore they have stated that though there have been many studies investigating the relationship

between dengue prevalence and rainfall it is hard to get such a relationship, especially in Malaysia.

A study has been conducted to compare the performance of regression models with two other models, Neural Network Model (NNM) and Hidden Markov Model (HMM) in dengue outbreak prediction [5]. The study states that regression models have less potential for over fitting and it is the widely used method over the other two in predictions. However, they require a considerable set of baseline data and are not very reliable at the extreme end of the range of the source data on which it is based.

Genetic Algorithms (GA) are heuristic search algorithms which have been used in improving classification algorithms and classifier systems [7]. GA s has been used to determine the relationship between climatic factors and dengue incidence trends in Singapore using time-series data [8]. Using GA rainfall and temperature has been identified as having a higher influence in dengue incidences trends. The experiment has been done with and without using GA. Without GA the maximum classification accuracy of an outbreak is 67% whereas applying GA it has obtained 77%-83% accuracy. One problem in applying GA is that there is a higher probability that it may not find the global optimum and retain in a local optima.

Serfling method is the standard CDC (Centers for Disease Control and Prevention) method used in flu detection in USA [9]. It is a cyclic regression model which uses the excess of flu mortality to determine an epidemic threshold. Using the excess of mortality as an index for identifying flu outbreaks is a well known method in many studies [9], [10], [11].

Auto-Regressive Integrated Moving Average (ARIMA) model is used in time series data analysis and is a generalization of the ARMA model [12]. ARIMA model has been used in predicting the number of dengue cases in Rio de Janeiro, Brazil [14] and in Southern Thailand [13]. In the former study it has been found out that number of dengue cases within a month can be predicted using one, two and twelve months prior data.

SVMs have been widely used in dengue outbreak predictions [8], [15]. Yusof *et al*. has explored using Least Squares Support Vector Machines (LS-SVM) in predictions. LS-SVM is a reformulation of the SVM algorithm and is faster in training process than standard SVM. Radial Basis Function (RBF) kernel has been used in training. Results are compared with the output of a Neural Network Model with the same input values. According to the study SVM has generated 86.64% prediction accuracy whereas NNM has generated only 65.58%. It states that the learning speed of SVM is very high compared to NNMs [15].

Using NNM in predicting the future dengue cases has been investigated [16]. A comparative study using HMM, ANN and RM has been done using the dengue case data from Selangor, Malaysia. The study clearly states that HMMs are rarely seen in dengue outbreak detection. The main reason behind this is HMM is a stochastic model rather than deterministic [5].

GIS has been used in many studies to map the spatial distribution of a dengue outbreak [4], [6], [17]. The relationship between dengue cases, population distribution, environmental impact and social economic factors affecting in spreading the epidemic has been investigated using GIS [6]. It has revealed a high relation between dengue incidences and population distribution in Johor Bahru district, Malaysia. Pathirana *et al*. has used GIS and remote sensing in risk mapping. Inverse Distance Weighted (IDW) interpolation has been used in converting points in to surfaces. The study has shown that high dengue incidences are normally associated with less abundance of rainfall.

III. DESIGN

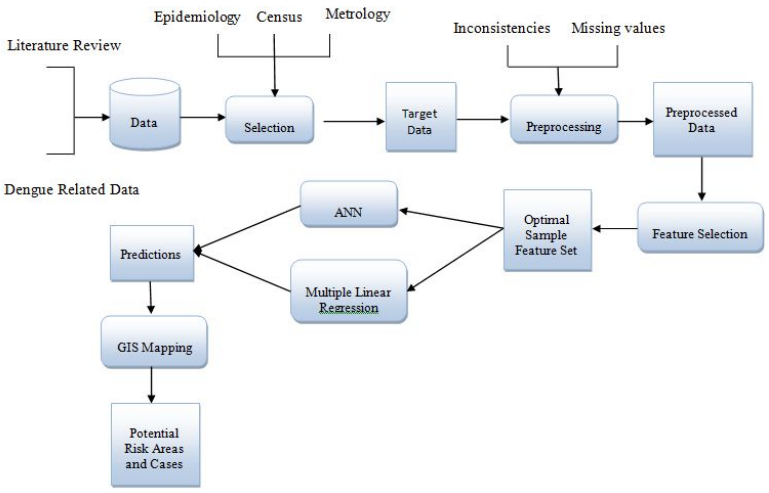

Fig. 1 The architecture of the proposed system

A. Hypothesis

The proposed model uses two hypotheses.

i. There are external factors related to causing a dengue epidemic.
ii. The relationship between risk factors and dengue outbreak patterns can be used to predict a future outbreak.

B. Data Gathering

Selecting an appropriate data set is highly important for the success of the project. The considered data set include

Epidemiological data
Metrological data
Census data

Weekly dengue fever and DHF data between 2010 and 2012 for Colombo and Matara districts in Sri Lanka were collected from the Epidemiology Unit, Colombo, Sri Lanka. Weekly and quarterly epidemiological reports from the epidemiology unit were used to gather conventional indices data. Dengue patients' data were collected from the offices of the Medical Officer of Health (MOH) and the Public Health Inspector (PHI).

Metrological data were gathered from the Metrology Department, Colombo, Sri Lanka and the Faculty of

Agriculture, University of Ruhuna, Sri Lanka. The Census data were acquired from the Census Reports of Sri Lanka.

C. Methodology

Design flow of the proposed model in the paper is depicted in Fig. 1. The design consists of six main phases namely data selection, preprocessing, feature selection, predictions using neural networks and multiple regression models, and GIS mapping of the results.

1. Data Selection

From the initial data set a target data set has been filtered out. Minimum, maximum, and average values of rainfall, temperature, and relative humidity were selected as metrological data. Conventional indices data and dengue cases data were considered for epidemiological data whereas population density is taken as census data.

2. Preprocessing

The target data set contains missing values, inconsistencies, redundant, and irrelevant values. It is essential to remove these inconsistencies, noise, and outliers. This preprocessed data set is fed in to the feature selection component.

3. Feature Extraction

The preprocessed data set contains relevant and irrelevant data both for predictions. We have to identify the factors which have an association with dengue outbreaks and how they are associated. A regression approach has been used for the purpose.

4. Non Linear Regression

A regression model is used in getting the relationship between risk factors and outbreaks. Separate regression formulas are generated to identify the relationship between climatic factors and population data with number of dengue cases. Whether a risk factor has a direct relationship with the outbreaks is analyzed based upon the correlation coefficient of each function. The optimal sample feature set contains the attributes which are having a higher correlation coefficient.

5. Outbreak Prediction

Predictions are done based on the attributes in the optimal sample feature set. Outbreak prediction is done using two methods separately and the outcomes of the two methods are analyzed to select the best method. Both methods output the number of outbreaks that will occur in a given period of time.

6. ANN

Prediction process of ANN can be divided in to three main stages, building the ANN, training the network, and testing the outcomes. When designing the network number of input nodes in the ANN is determined by the number of attributes in the optimal sample data set. The number of hidden layers and the number of neurons in each hidden layer has a major impact on the target output set. A trial and error method has been used in determining the number of hidden layers and neurons. The output parameter is the number of dengue cases that may occur for that specific area within a given time.

Forward propagation and backward propagation methods are used in training and learning process.

7. Multiple Linear Regression

In the regression dengue case data is used as the dependent variable. Predictor variables were chosen from the optimal sample feature set. A regression formula consisting of these attributes is generated that can be used for predictions when a new input data set is given. Root Mean Square Error method is used in measuring the model performance.

8. GIS Mapping

Spread of the disease within a population can be addressed using spatial analysis tools. Thus the predictions done using ANN and regression models are mapped in to a spatial distribution using GIS.

IV. Implementation

This section mainly focuses on implementation details of the prediction model, comparing the methodologies used.

A. Feature Selection

To identify the relevant factors which have an impact in causing outbreaks, regression models have been created using the statistical tools. First we have tried to create linear regression models for each data set where the relationship between two variables is modeled by fitting a linear equation. However, none of the factors in the sample data set have shown a considerably high correlation which implies that there is no linear relationship between the original factors with number of cases. Here we have considered the r-squared value and the adjusted r-squared value to determine relationship between the modeled variables. An adjusted r-squared value which is greater than 50% is considered to indicate a higher correlation between the variables. Fig. 2 shows the regression model obtained for average rainfall and dengue cases for Matara district which clearly shows that linear regression is not applicable.

The next step taken is to transform the data to make it more linear. Transforming a variable means that using a mathematical operation to change its measurement scale. After applying mathematical models we have been able to obtain very high adjusted r-squared values for the relationships between variables. For example Fig. 3 shows the relationship between dengue cases and rainfall using the power model stated above and the adjusted r-squared value is very high in this model, which is 94.5%.

Using the regression approach stated, the optimal sample feature set has been identified which consists of eight original factors namely average rainfall, maximum rainfall, maximum temperature, minimum temperature, average temperature, average relative humidity, maximum relative humidity and Bretau index value for each week.

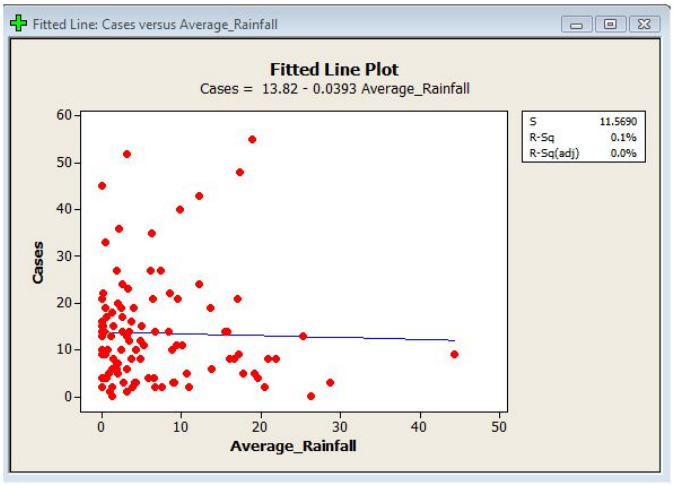

Fig. 2 Linear regression model for average rainfall and dengue cases

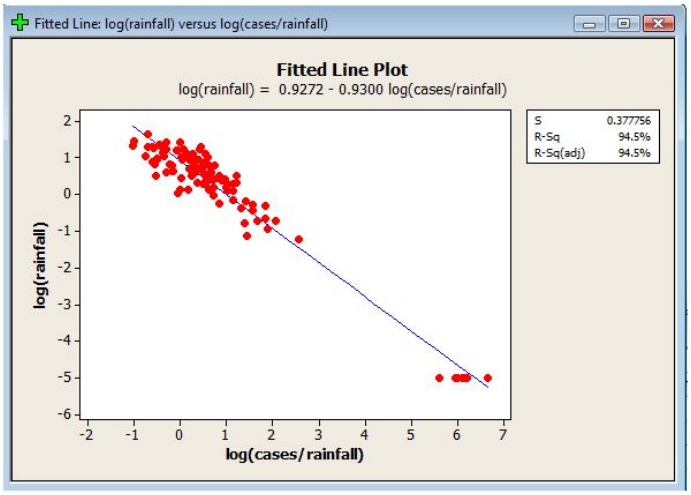

Fig. 3 Power model for average rainfall and dengue cases

B. Prediction Model

We have implemented a neural network to predict the number of future outbreaks, number of deaths, intensity of the disease for a given location and a time period. As climatic and epidemiological factors differ in district wise we have created two separate neural networks for the two districts.

The prediction process is divided in to three main steps.

Building the NN

Training/Learning

Testing/Validating

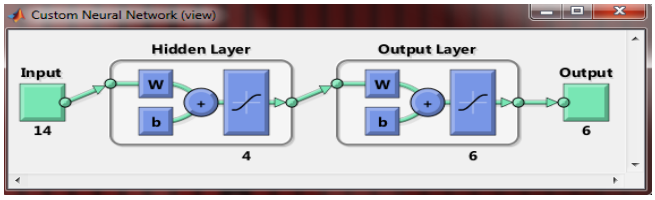

Fig. 4 Proposed neural network model for Matara district

V. Evaluation

Initial experiments and testing has been restricted only for Colombo and Matara districts data. Data in 2010 and 2011 was used for training purposes and data in 2012 was used for testing.

A. Feature Extraction Unit

Separate regression models have been created for each feature in the data set to identify the optimal sample feature set. These models are evaluated separately to identify their influence in causing an outbreak and to test their influence for the pattern recognition model.

TABLE I
REGRESSION MODELS FOR MATARA DISTRICT

Feature	Regression Equation	R-Squared	Adjusted R-Squared
Maximum Rainfall	$cases = 12.7 \times (\max\_rainf all)^{-0.063}$	95.2%	95.2%
Average Rainfall	$cases = 9.91 \times (average\_rainf all)^{-0.075}$	94.5%	94.5%
Maximum Relative Humidity	$cases = 493 \times 10^5 \times (\max\_rh)^{-3.49}$	51%	47%
Average Relative Humidity	$cases = 101 \times 10^5 \times (average\_rh)^{-3.75}$	62.3%	61.9%
Minimum Relative Humidity	$cases = 168 \times 10^4 \times (\min\_rh)^{-3.582}$	61.4%	61%
Maximum Temperature	$cases = 0.03137 / (1 - 0.6627 \log(\max\_temp))$	98%	96%
Average Temperature	$cases = 0.03802 / (1 - 0.6857 \log(average\_temp))$	98%	98%
Minimum Temperature	$cases = 0.2165 / (1 - 0.706 \log(\min\_temp))$	92%	92%

According to the results obtained higher number of dengue cases is associated with less abundance of rainfall and relative humidity.

cases α (max_rainfall)$^{-0.063}$

cases α (average_rh)$^{-3.78}$

This means that number of dengue cases is high in less rainy periods. According to the epidemiological reports dengue mosquitoes start breeding at the end of a rainy season and there is a lag time between occurring outbreaks. The obtained results confirm this factor.

According to Table I higher number of dengue cases is associated with high temperature. This is because egg hatching occurs within an optimal temperature range.

Table II indicates the regression models obtained for Colombo district data. Unlike for Matara district, results do not show a higher correlation with outbreaks. This may be due to the fact that Colombo is a highly urbanized area and external factors affect causing outbreaks. External factors such as migration rate, sanitation facilities, health care facilities, pollution, and poverty may affect the outcome of the model.

TABLE II
REGRESSION MODELS FOR COLOMBO DISTRICT

Feature	Regression Equation	Adjusted R-Squared
Maximum Rainfall	$cases = 887 \times (\max\_ra\inf all)^{-1.303}$	34.2%
Average Rainfall	$cases = (average\_ra\inf all)^{3.344} / 209.9$	48%
Maximum Temperature	$cases = 0.3226 / (1 - 0.6724\log(\max\_temp))$	98.6%
Minimum Temperature	$cases = 1.496 / (1 - 0.704\log(\min\_temp))$	98%

B. Neural Network Model

Performance of the neural network has been analyzed for different number of hidden nodes. As the data set is not very large only one hidden layer has been used changing the number of hidden nodes.

TABLE III
NUMBER OF HIDDEN NODES

Equation for Hidden Nodes	Number of Hidden Nodes	Classification Accuracy
2n + 1	29	83%
$(n+m)/2 + n^{1/2}$	13	92%
n/2	7	81%

In the neural network model training accuracy has been high giving 83.6% for Matara and 89.5% for Colombo district respectively indicating a good fitting for the model.

The testing set accuracy has been checked for the model. Accuracy for Matara district is 62% and for Colombo district it is 67%. There is a rapid decrease in the testing accuracy compared to training accuracy. This may be due to the reason that the training set is not very large. The training data set is not equally divided and some classes contain very few data.

VI. CONCLUSION

Many investigations and researches have been done to model Early Warning Systems (EWSs) for different infectious diseases based on risk factors. Currently such EWSs are used for predictions in developed countries. However, in Sri Lanka no such systems are currently used for disease planning. Control measures can be taken if such predictions are made prior to an epidemic. This research was intended to address the above mentioned problems and to model an Early Warning System (EWS) for dengue outbreak prediction.

Results obtained for Matara district shows that higher incidences of outbreaks are related with less abundance of rainfall and relative humidity and with higher temperature. However, the results of more urbanized areas (Colombo) are more complex, and most of the factors do not show a direct correlation with outbreaks. Power model for relative humidity shows a higher correlation indicating an inverse relationship but other factors do not show much influence. This may be due to the various influences in urbanized areas such as migration rates, pollution rates, immunity of the population etc.

For outbreak prediction two separate neural networks were implemented for Colombo and Matara districts. Outputs were classified in to six (6) classes. The system had some problems while classifying because the data set was not enough to divide data equally for all the classes. According to the results, an outbreak was predicted with 86.5% accuracy in Matara district and 89.5% accuracy in Colombo district. However, the testing accuracy is low due to the unavailability of a substantial amount of data.

The performance of the neural network is considerably high compared to that of the multiple regression model. Thus, for predictions, neural network model is more suitable for Colombo and Matara districts.

REFERENCES

[1] World Health Organization. http://www.who.int/csr/disease/dengue/en/.
[2] Epidemiology Unit,Ministry of Health,Sri Lanka. http://www.epid.gov.lk/web/index.php?option=c articleid = 171Itemid = 487lang = en.
[3] http://www.epid.gov.lk/web/index.php?option=com_content&view=article&id=171&Itemid=487&lang=en.
[4] S. Pathirana, M. Kawabata and R. Goonatilake, "Study of potential risk of dengue disease outbreak in Sri Lanka using GIS and statistical modeling," JRuralTropPublicHealth, VOL 8, p. 8-17, May 2009.
[5] N.A. Husin, N. Salim and A.R. Ahmad, "Simulation of dengue outbreak prediction," proceedings of the postgraduate annual research seminar, 2006.
[6] S.B. Seng, A.K. Chong and A. Moore, "Geostatistical modelling, analysis and mapping of epidemiology of dengue fever in Johor State, Malaysia," presented at SIRC, 2005.
[7] D. Whitely, "A genetic algorithm tutorial," Computer Science Department, Colorado State University.
[8] X. Fu, C. Liew, H. Soh, G. Lee, T. Hung and L.C. Ng, "Time-Series infectious disease data analysis using SVM and genetic algorithm," IEEE Congress on Evolutionary Computing, 2007.
[9] D.J. Muscatello, P.M. Morton, I.Evans and R.Gilmour, "Prospective surveillance of excess mortality due to influenza in New South Wales: feasibility and statistical approach", December 2008.
[10] R.E.Serfling, "Methods for current statistical analysis of excess pneumonia-inluenza deaths", 1963.
[11] S.D.Collins and J.L.Lehman, "Influenza epidemics during 1951-56 with a review of trends".
[12] A. Moore, G. Cooper, R. Tsui and M.Wagner, "Summary of biosurveillance-relavent statistical and data mining technologies," February 2002.
[13] Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA models, S. Promprou, M. Jaroensutasinee and K. Jaroensutasinee, Dengue Bulletin, volume 30, 2006.
[14] Time series analysis of dengue incidence in Rio de Janeiro, Brazil, P.M. Luz, B.V.M. Mendes, C.T. Codeco, C.J. Struchiner and A.P. Galvani, Am.J. Trop.Med.Hyg, 79(6), 2008, pp 933-939.
[15] Y. Yusof and Z. Mustaffa, "Dengue outbreak prediction: A least squares support vector machines approach," International Journal of Computer Theory and Engineering, Vol. 3, No 4, August 2011.
[16] B.G. Cetiner, M. Sari and H.M.Aburas, "Recognition of dengue disease patterns using artificial neural networks," 5th International Advanced Technologies Symposium, May 2009.
[17] P. Jeefoo, N.K. Tripathi and M. Souris,"Spatio-Temporal diffusion pattern and hotspot detection of dengue in Chachoengsao province, Thailand, Interanational Journal of environmental research and public health, 2011.