A Novel Approach towards Prediction of Mosquito-Borne Diseases

Gresha Bhatia
Deputy Head of Department, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, Maharashtra, India
gresha.bhatia@ves.ac.in

Shravan Bhat
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, Maharashtra, India
2017.shravan.bhat@ves.ac.in

Vivek Choudhary
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, Maharashtra, India
2017.vivek.choudhary@ves.ac.in

Aditya Deopurkar
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, Maharashtra, India
2017.aditya.deopurkar@ves.ac.in

Sahil Talreja
Student, Department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, Maharashtra, India
2017.sahil.t@ves.ac.in

***Abstract*—Communicable diseases, particularly vector-borne diseases, are a leading cause of morbidity and mortality worldwide. In the age of big data, answering broad-scale, basic issues regarding the nuanced nature of these diseases would increasingly necessitate the synthesis of diverse datasets to produce new biological information. Data mining and deep learning have the potential to make important advances in understanding fundamental aspects of vector-host-pathogen interactions, and their use in this area should be welcomed. Data mining and machine learning approaches such as deep learning lag in this area and should be used in conjunction with existing methods to speed hypothesis and information generation. Deep learning is used in this research to forecast mosquito-borne diseases based on environmental factors.**

Keywords—vector-borne diseases, data mining, deep learning, environmental factors

I. INTRODUCTION

Machine learning is the process of programming computers to improve their output based on previous data or examples[1]. The study of computer systems that learn from data and experience is known as machine learning. There are two passes in the machine learning algorithm: Training and testing. The applications of machine learning include speech recognition, spam filtering, etc. One such implementation in the field of healthcare is prediction of diseases.

Climate change is widely acknowledged to play a role in the spread of a variety of infectious diseases, some of which are among the leading causes of mortality and morbidity in developing countries. These diseases often manifest as epidemics, which may be caused by climatic changes that favour higher transmission rates. With the rising demand for operational disease early warning systems (EWS), recent developments in the availability of climate and environmental data, as well as increased use of geographic information systems (GIS) and remote sensing, make climate-based EWS more technically feasible. Prediction of a disease by using climate data and previous history machine learning technology is facing few struggles from the past years[2]. Machine Learning technology offers a strong forum in the medical sector for efficiently addressing healthcare issues.

Human illnesses caused by parasites, viruses, and bacteria spread by vectors are known as vector-borne diseases. Malaria, dengue fever, schistosomiasis, human African trypanosomiasis, leishmaniasis, Chagas disease, yellow fever, Japanese encephalitis, and onchocerciasis kill over 700,000 people per year[3].

In general, temperature, rainfall, and humidity have a significant impact on the geographic distribution and growth rate of insect vectors. Temperature increases the metabolic rate of mosquitos, increases egg production, and increases the frequency of blood feeding. Rainfall has a major effect as well, but it is more difficult to predict. Rainfall has an indirect impact on vector longevity due to its effect on humidity; relatively wet conditions can establish favourable insect habitats, resulting in an increase in disease vector geographical distribution and seasonal abundance. In other situations, if flooding washes away breeding grounds, heavy rainfall may have devastating implications for local vector populations[4-6]. As a result, the study aims to use a time series analysis model to forecast mosquito-borne disease incidence using weather (i.e., precipitation, humidity, rainfall, and temperature) as input factors.

II. METHODOLOGY

Identifying and estimating the number of people afflicted with mosquito-borne disease using the training dataset consisting of malaria, dengue, chikungunya, zika, yellow fever cases along with corresponding meteorological factors and adding the pattern to the research dataset. Fig. 1 shows the system design for our prediction model which comprises of Data Collection Unit on which Data Analysis is performed, Data Preprocessing Unit where mainly filtering and cleaning are done on the gathered data to give better outcomes, Deep Learning models which were namely LSTM, GRU and RNN for making future predictions using the preprocessed data and finally Visualization dashboard

where these results are shown in tabular and graphical formats.

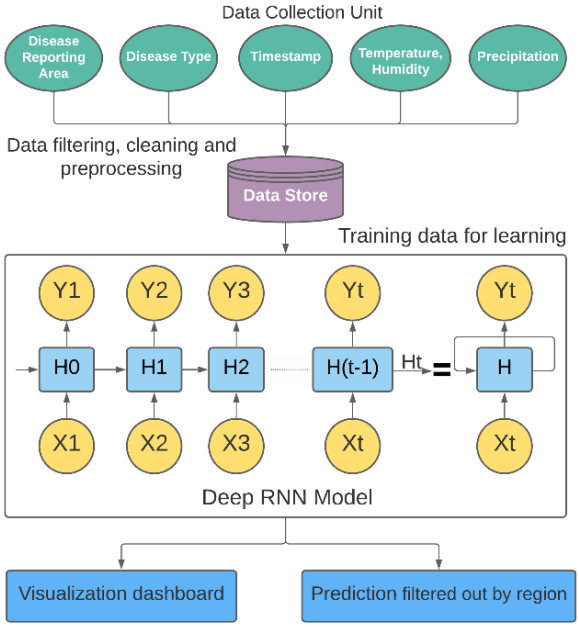

Fig. 1. System Design

A. *Data Collection*

Two datasets were used for the model.

- San Juan and Iquitos Dengue dataset provided by Drivendata[7-8] via Centers for Disease Control and Prevention[9] database which is the national public health institute in the United States of America.
  San Juan: 1990-2008
  Iquitos: 2000-2010
- Mumbai region 24 wards Malaria, Dengue and Chikungunya dataset collected from various hospitals for 2016.

The main factors considered for the prediction model were Disease reporting area, Disease name, Timestamp (year, weekofyear, week_start_date), Temperature, Precipitation, Humidity, Rainfall.

B. *Data Analysis*

The data collected was explored to check for any easily visible relationship between themselves and for outliers. For this correlation heat matrix, week wise column data plotting and density plotting of the data was done.

C. *Data Preprocessing*

Before feeding the data into the Prediction model, following data cleaning and preprocessing steps are performed

- Null values were checked and filled using the forward fill method.
- Data converted into weekly data.
- Data standardized using mean and standard deviation.
- Dataset splitting into training, validation, and testing sets.

D. *Building Model*

Deep RNN model was used for the prediction as it is well suited for sequential data, time series prediction. The model used here is LSTM as they have the promise of being able to learn the context required to make predictions while also being able to preserve long term dependencies within data.

E. *Visual Representation*

To display the spread of a disease in a more user-friendly manner, the following visualization patterns were used

- Mosquito borne disease region wise status which shows region name, mosquito disease type, cases and total deaths in that region.
- Disease heatmap is used to show how severe the disease is in a particular region.
- Disease current and prediction timeline map to display the spread of a disease region-to-region wise.

III. Proposed System

The heart of the disease prediction model is the Deep Recurrent Neural Network (RNN) which predicts the possibility of a mosquito borne disease in a particular region.

A. *Simple Recurrent Neural Network (RNN)*

Simple RNN[10] are a type of neural network where the output from the previous step is fed as input to the current step. RNN converts the independent activations into dependent activation by using the same parameters for each input and hidden layer and has a memory which remembers all the calculations. Simple RNN suffers from the problem of vanishing gradients where the gradients carry information and when the gradients become too small there is no change in the output improvement.

Formula —

$$h_t = f(h_{t-1}, x_t) \quad (1)$$

h_t = current state (2)

h_{t-1} = previous state (3)

x_t = input state (4)

B. *Long Short-Term Memory (LSTM)*

Long Short-Term Memory (LSTM)[11] is used to solve the problem of vanishing gradients and it is capable of learning long term sequences by remembering the information for a longer time. LSTM has three gates — Forget gate which decides how much past data it should remember, Update / Input gate which decides how much this units add input to the current state and Output gate which decides which part of the current cell makes it to the output.

Forget gate —

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \quad (5)$$

Update / Input gate —

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\acute{C}_t = tanh(W_C . [h_{t-1}, x_t] + b_C) \quad (7)$$

Output gate —

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t * tanh(C_t) \quad (9)$$

C. Gated Recurrent Unit (GRU)

Gated Recurrent Unit[12] is variation of LSTM unit as both the units are similar in design and produces equal results in some cases. GRU solves the problem of vanishing gradient by using two gates - update and reset gates. The update gate is used to determine how much past information needs to be passed in the future and reset gate determines how much of past information to forget. GRU is a fast and compact model, and it requires fewer parameters for execution, so less memory is required and also faster execution than LSTM. GRU units control the flow of information without using cell memory units.

Update gate —

$$z_t = \sigma(W^{(z)} x_t + U^{(z)} h_{t-1}) \quad (10)$$

Reset gate —

$$r_t = \sigma(W^{(r)} x_t + U^{(r)} h_{t-1}) \quad (11)$$

Current memory content —

$$h_t = tanh\,(W x_t + r_t \odot U h_{t-1}) \quad (12)$$

Final memory at current timestep —

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h_t \quad (13)$$

Evaluation parameter used are —

- Mean Absolute Error: Computes the mean absolute error between the labels and predictions[13].
- Mean Squared Error: Computes the mean squared error between labels and predictions[14].
- Accuracy of the model.

Disease prediction model should give output as —

- Which cities have high probabilities of spreading of a specific type of Mosquito borne disease.
- Expected number of people that might get infected by a particular disease in that region.

IV. Results

A website implementation for the proposed system was done.

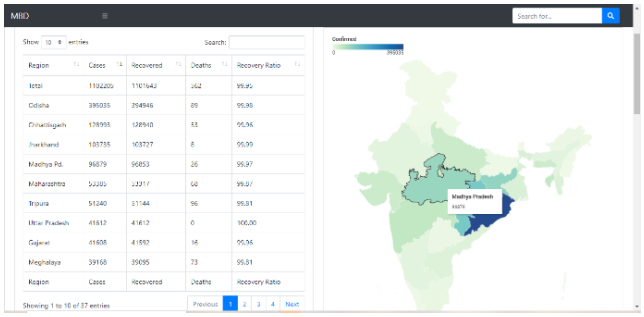

Fig. 2. Malaria in India

Fig. 2 shows the number of Malaria Cases and deaths in India. This is done using tabular form and in choropleth mapping.

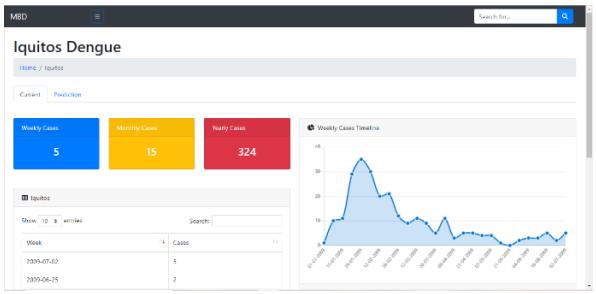

Fig. 3. Current Cases in a Region

Fig. 3 shows the present number of cases for the dengue in Iquitos region using tabular form, cards and a timeline map.

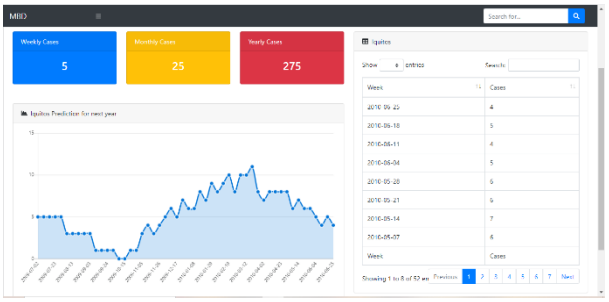

Fig. 4. Prediction in a Region

Fig. 4 shows the predicted number of cases for dengue in Iquitos region using tabular form, cards, and a timeline map.

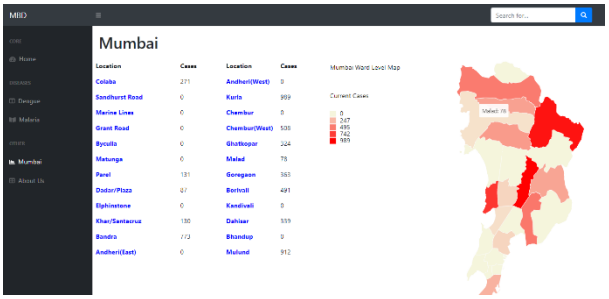

Fig. 5. Cases in Mumbai

Fig. 5 shows the representation of ward wise number of cases for Malaria, Dengue and Chikungunya in Mumbai for 2016 data using tabular form and choropleth map.

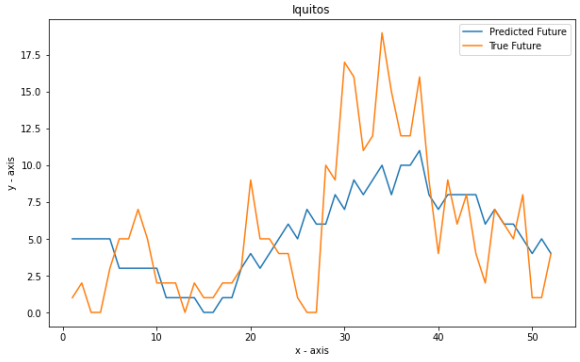

Fig. 6. Dengue Prediction

Fig. 6 shows data for the Iquitos region. This figure shows the true future in red and prediction in blue for the number of dengue cases for 52 weeks. Here X-axis represents weeks while Y-axis represents the number of cases for that week.

Mean Absolute Error = 13.05

Yearly Accuracy = 92%

The LSTM based model was then run on the second dataset consisting of different sectors of Mumbai. Fig. 7 and Fig. 8 shows the Model Accuracy and Model Loss obtained for the same, respectively.

Final Loss (Mean Squared Error): 1.8473

Final accuracy: 0.8444%

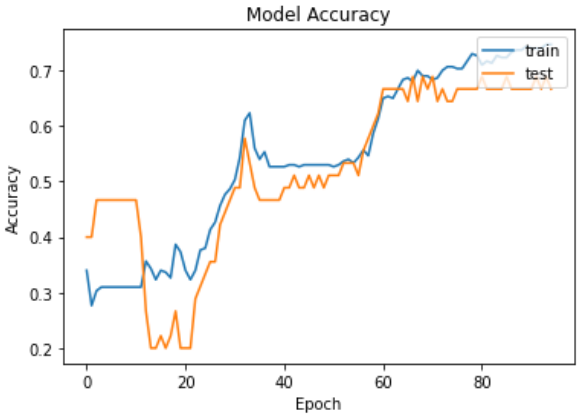

Fig. 7. Model Accuracy

Fig. 7 gives accuracy of model vs number of epochs for Mumbai region dataset. The blue and orange line represent training and testing part accuracy, respectively.

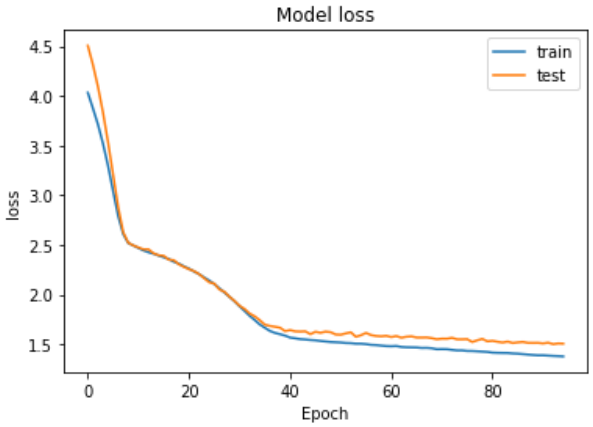

Fig. 8. Model Loss

Fig. 8 gives loss of model vs number of epochs for Mumbai region dataset. The blue and orange line represent training and testing part model loss, respectively.

Comparison between LSTM, GRU and Simple RNN model loss and accuracy.

TABLE I. MODEL LOSS AND ACCURACY

Model	Loss	Accuracy
LSTM based	1.84	**0.84**
GRU based	**1.81**	0.82
Simple RNN	3.06	0.26

GRU requires less training parameters, which means it consumes less memory, performs quicker, and trains faster than LSTM whereas LSTM model being more sophisticated at the same time more complex is being more accurate on a dataset using a longer sequence.

V. CONCLUSION AND FUTURE SCOPE

In this paper we have developed predictive models for mosquito-borne diseases. The steps involved were Data Collection, Data Analysis, Data Preprocessing, Building Model and Visual Representation. We have used Deep RNN for training the model with historical climatic and disease data to make future predictions. The accuracy and loss obtained for longer time series have yielded significantly better results as seen from the results obtained from Iquitos, San Juan dataset and Mumbai region dataset.

In this paper, we investigated the principles of data mining and deep learning as they relate to vector-host-pathogen prediction. A potential increase in deep learning applications in the field may be beneficial, especially when coupled with other approaches such as feature engineering, cross-validation, model selection, and crowdsourcing. While the application of such approaches to specific problems in vector-borne diseases is still in its infancy and faces many expected obstacles, these strategies have considerable potential and should be encouraged to apply new ideas to old problems as large, complex systems biology databases become accessible in the field.

REFERENCES

[1] T. M. Mitchell, “Machine learning WCB”: McGraw-Hill Boston, MA:, 1997.
[2] Katrin Kuhn, Diarmid Campbell-Lendrum, Andy Haines, Jonathan Cox ‘Using Climate to Predict Infectious Disease Outbreaks: A Review’ (WHO/SDE/OEH/04.01) https://www.who.int/globalchange/publications/en/oeh0401.pdf
[3] Vector Borne Diseases https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases
[4] Müller R., Reuss F., Kendrovski V., Montag D. (2019) Vector-Borne Diseases. In: Marselle M., Stadler J., Korn H., Irvine K., Bonn A. (eds) Biodiversity and Health in the Face of Climate Change. Springer, Cham. https://doi.org/10.1007/978-3-030-02318-8_4
[5] Ahmed, T., Hyder, M. Z., Liaqat, I., & Scholz, M. (2019). Climatic Conditions: Conventional and Nanotechnology-Based Methods for the Control of *Mosquito* Vectors Causing Human Health Issues. *International journal of environmental research and public health*, *16*(17), 3165. https://doi.org/10.3390/ijerph16173165
[6] Fouque, F., Reeder, J.C. Impact of past and on-going changes on climate and weather on vector-borne diseases transmission: a look at the evidence. *Infect Dis Poverty* 8, 51 (2019). https://doi.org/10.1186/s40249-019-0565-1
[7] https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/
[8] Peter Bull, Isaac Slavitt, Greg Lipstein. (2016). Harnessing the Power of the Crowd to Increase Capacity for Data Science in the Social Sector; arXiv:1606.07781v1 [cs.HC]
[9] https://www.cdc.gov/dengue/statistics-maps/index.html
[10] https://www.tensorflow.org/guide/keras/rnn
[11] https://keras.io/api/layers/recurrent_layers/lstm/
[12] https://www.tensorflow.org/api_docs/python/tf/keras/layers/GRU
[13] https://www.tensorflow.org/api_docs/python/tf/keras/losses/MeanAbsoluteError
[14] https://www.tensorflow.org/api_docs/python/tf/keras/losses/MSE