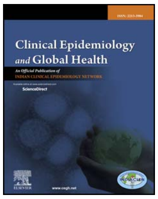
Clinical Epidemiology and Global Health 7 (2019) 121–126

Contents lists available at ScienceDirect

Clinical Epidemiology and Global Health

journal homepage: www.elsevier.com/locate/cegh



Artificial neural network based prediction of malaria abundances using big data: A knowledge capturing approach

Thakur Santosh, Dharavath Ramesh*

Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad – 826004, Jharkhand, India

ARTICLE INFO

Keywords:
Malaria prediction
Primary health centers (PHCs)
Big data
Artificial neural networks (ANNs)

ABSTRACT

Background and objective: Malaria is one of the most prevalent diseases in urban areas. Malaria flourishes in sub-tropical countries and affect the public health. The impact is very high, where health monitoring facilities are very limited. To minimize the impact of malaria population in sub-tropical domains, a suitable disease prediction model is required. The objective of this study is to determine the malaria abundances using clinical and environmental variables with Big Data on the geographical location of Khammam district, Telanagana, India.

Methods: Prediction model is based on the data collected from primary health centres of department of vector borne diseases (DVBD) of Khammam district and satellite data such as *rain fall, relative humidity, temperature and vegetation* taken for the time period of 1995–2014. In this study, we test the efficacy of the artificial neural network (ANN) for mosquito abundance prediction. Prediction model was developed for the period of 2015 using a feed forward neural network and compared with the observed values.

Results and conclusions: The results vary from area to area based on clinical variables and rainfall in the prediction model corresponding to areas. The average error of the prediction model ranges from 18% to 117%. Clinical data such as number of patients treated with symptoms and without symptoms can improve the prediction level when combined with environmental variables. We perform preliminary findings of malaria abundances by collecting clinical big data across different seasons. Further, more exploration is required in prediction of malaria using big data to improve the accuracy in real practice. In this manuscript, we perform some preliminary findings of malaria abundances by collecting larger data across different seasons. Till today, many models have been developed to examine the malaria prediction with different approaches, but malaria prediction with environmental and clinical data is a new approach with big data analysis.

1. Introduction

There have been growing instances about the effects of global warming, including the growth and activity of insects which cause the disease to humans.[27] To predict malaria abundances, different mechanistic models have been developed. Malaria prediction models were started from early works of Christopher[1] in 1911 and till today, many models have been developed in endemic countries. Each work has a different interpretation and variety of prediction. Commonly, these methods use data on environmental conditions to forecast malaria for a certain period of time.[2] Some researchers have developed different models based on chemical property instances, such as rainfall and temperature variations.[12,28,29,32] But, in previous works, predictions based on symptomatic clinical conditions, asymptomatic clinical conditions, and environmental factors in terms of big data approach have not been explored much to provide suitable clinical treatments. Among the popular approaches, the artificial neural networks (ANNs) were widely applied to develop good predicting models, especially in the fields of Bioinformatics[30] and environmental studies.[31] The instance of asymptomatic carriers does not require treatment, but they act as reservoirs for the parasite.[3] An improper anti-malarial treatment and asymptomatic malarial case has the high potential of prediction where, various environmental factors support for the growth of the malaria parasite.[33]

In South Asia, India is one of the endemic place of malaria. In India, the eradication of malaria has started in 1950s. Even every year, around one million cases were noted and 1000–1500 deaths were reported.[4,5] Malaria is considered as one of the major disease and has got the spreading sequences, especially in the states of southeast India includes Orissa, Chattisgarh, Telangana, and Andhra Pradesh. We conduct our study for malaria abundances in the geographical location of the Khammam district area which is situated in the state of Telangana on the bank of the Godavari river across the borders of Orissa and Chhattisgarh states with latitude 37′18 *N* and longitude 18 *E*. The

* Corresponding author.
E-mail addresses: santosh.t68@gmail.com (S. Thakur), drramesh@iitism.ac.in (R. Dharavath).

https://doi.org/10.1016/j.cegh.2018.03.001
Received 2 October 2017; Received in revised form 9 January 2018; Accepted 16 January 2018
Available online 08 March 2018

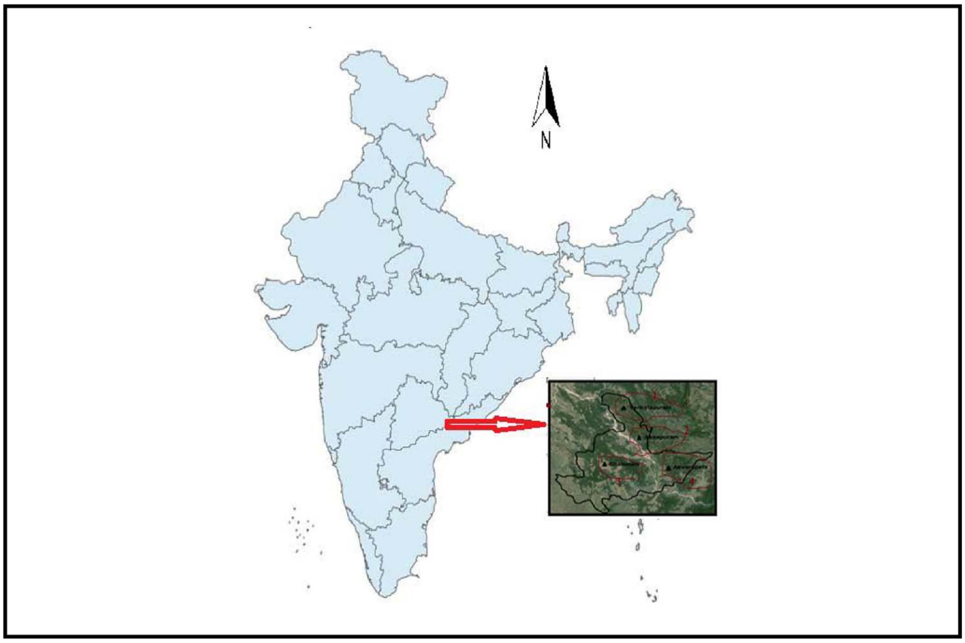


Fig. 1. Map of India showing areas in Khammam District, selected for study.

related malaria abundances were monitored at the sites of (i) Venkatapuram, (ii) Aswapuram, (iii) Khammam, and (iv) Aswaraopeta. The geographical location of individual site is shown in Fig. 1. Khammam district is sub-tropical climate with wet and warm humid weather and densest moist forest creates favorable conditions of malaria.[6]

2. Methods

2.1. Big data

In the recent era, human society is witnessing in generating of huge data from different sources in different formats. Different technologies came in practice to analyze the data with different methodologies and making the data in human readable format. The combination of these trends is termed as Big data. Big data in health care are used to predict the epidemics and avoid curable deaths.[7] In disease monitoring, at every time, we store the data and this leads to big data and this makes a vital piece of time-based public observation.[13] Using big data, we can analyze disease patterns and able to track the outbreaks of contagious disease which will help to speed up the response against outbreaks.[8] In recent years, popularity has been gained in health care research with big data for early disease diagnosis and optimal management.[9,10,11,25]

With the advances in healthcare systems, huge data are generated and stored in various locations. With the help of these systems, public health predictive analyses can be made.[18] The hospitals of the United States are using predictive analysis from past patient records to take decisions and to further optimize the effect for better understanding of diseases.[19] Australia has made apps for health insurance based on Big data analysis to detect errors and fraud for better patient services.[20] Different researchers have developed different machine learning algorithms and employed to predict asthma using big data. This has been formulated by analyzing Clinical data, Twitter data, Google data, and Sensor data for air pollution and developed different prediction models.[34,35] In terms of predictive analysis, recent initiatives in Big data for health care are grown up drastically with the rate of 40%.[23,24,26,36]

2.1.1. Clinical data

Related datasets are collected from Primary Health Centers (PHCs) of all geographical sites. These PHCs are used to collect the samples from endemic places of these sites to monitor the malaria abundances in the geographical location of Khammam district. In this geographical location, in total, 52 primary health centres (PHCs) were established by the government of Telangana, out of which 32 PHCs were highly endemic. All PHCs staff is highly trained in malaria testing, diagnosis, and treatment. In this study, we consider environmental factors such as RH (relative humidity), temperature, rainfall and clinical factors with the number of positive cases without symptoms (ASC) to predict malaria abundances. The data characteristics of each primary health center for a time period between Jan 1995–December 2014 is shown in Table 1.

2.1.2. Environmental data

This study uses the environment data derived from satellite sensors called TRMM and MODIS on Terra satellite. The TRMM product (TRMM3B42) uses for daily rainfall and estimates at a spatial resolution of 0.25° * 0.25° (27.8 km * 27.8 km).The MODIS product(MOD11A2) estimates day time and the night time temperatures (LST) using 8-day composite image at 1km*1 km spatial resolution. A 16-day composite processed images with an enhanced vegetative index (EVI) has been obtained from MODIS product (MOD13A1) at a resolution of 0.5*0.5. The meteorological survey data acquisition center (MOSDAC) provides daily RH(relative humidity) at 2 m and 8 m height.

Table 1
Data Characteristics of each primary health center (PHC).

	Total No. of cases	Average age (years)	%male	%female	Average Temperature (°C)	Average Rainfall(mm)
Khammam	1152	11.6	45	55	27.4	1.66
Aswapuram	2462	14.2	38	62	29.1	1.5
Aswaraopeta	14783	14.2	48	52	28.4	1.33
Venkatapuram	5284	19.1	56	46	27.9	1.22

2.2. Artificial neural network

Artificial neural networks (ANNs) are divided into different types. The ANN focuses on different applications such as weather forecasting, data mining, traffic, image processing, Big data etc. with current research attractions.[13] The advantage of ANNs model is that it can capture both linear and nonlinear data and produce results with a high degree of accuracy.[14,15] In a recent work, a prediction model is presented to predict mosquito abundances by using the simple back propagation method.[16] A feed forward neural network model with a single hidden layer provides time series modelling which is widely used in forecasting.[17] It contains input nodes, hidden nodes and output nodes. Here, in order to investigate the competence of ANN process, we attach *weight* to determine the corresponding neuron and data pass signals between neurons. These signals are processed as an integrated function which combines the signals and proceeds for activation function and finally passes the output. This model was developed to forecast the values for every two months. As a result, it has provided over one year of prediction based on the past values given.

2.3. Measurement

To predict malaria abundances, we consider meteorological variables as input. In this, we use meteorological data, including Temperature(T), Rainfall (RF), Relative humidity (RH), Asymptotic cases (ASC), and Vegetative Index (VI) as input variables (Fig. 2). In our model, we collect data as average daytime temperature, the night time temperature, average rain fall of each day in mm, and average EVI at one-kilometre resolution and clinical data of each day. The data are aggregated in a monthly frequency form. Approximately, overall 12% of the observations were missing for environmental data and 7% of clinical data. To handle the missing data, we adopted multivariate imputation by chained equations in data pre-processing using MICE R package. The time series of the environmental variable were created once in a month for each area and averaged. The environmental and clinical variables begin from Jan 1995 to Dec 2014 at each geographical area as shown in Fig. 1.

The finest combination of environmental and clinical covariates have been determined by using training data. The training data were reassessed with the original values. After that, the values were adjusted according to the model. The substantial lags between the response series and predictors series were considered for model building. The working strategy of the model was studied through an auto correlation function(ACF), and partial auto correlation function (PACF). In this study, a neural model has been used to generate monthly forecasts for 52 weeks. The visualization of our model is depicted in Fig. 2.

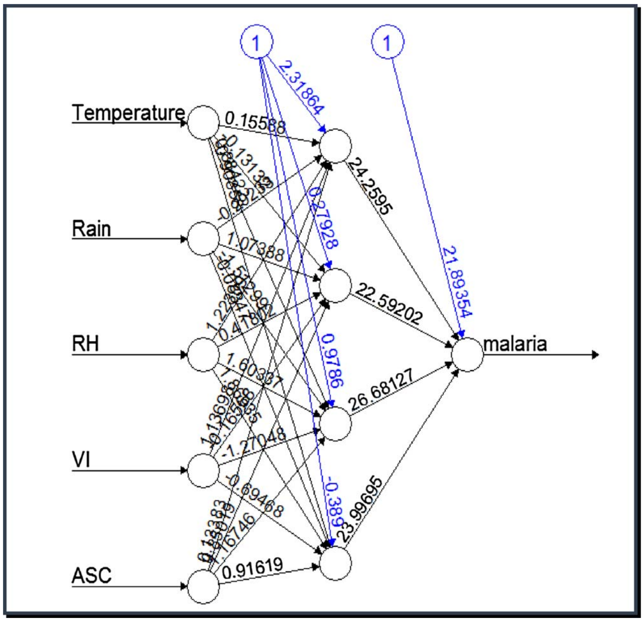


Fig. 2. Visualisation of feed forward neural networks.

Based on the forecast data, we have extracted and predicted relative malaria abundances from January 2015 to December 2015. We train the neural network and neural model to capture the data from 1995 to 2014 and tested from January 2015 to December 2015. In order to predict the suitable information without the noise, training data are used for building the model and test data are used for predicting the accuracy. The package of neural net uses the same type of function of the neural network. The neural network processes the output depending upon the given input. If the training instance is not completed properly, the predicted output will differ from the actual output. The root mean square percentage error (RMSPE) function quantifies the error which has been occurring between the actual output and predicted output. The root mean square percentage error(RMSPE) is used to measure the error in the following manner.

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=1}^{n}(x_i - y_i)^2}*100 \tag{3}$$

Where x_i is the observed value of the area and y_i is the predicted value and n be the number of predictions or observations in the corresponding geographical area.[24,26] There are multiple predictions occurred for each geographical area. For each individual geographical area root mean square percentage error is calculated. The value of the bias is set to zero and upper bound of *RMSPE* is set to 150%. Finally, the observed cases were compared with predicted cases. The ANN model used in this study is a feed-forward network, which is implemented in *R* studio (*version* 3.1.2) with neural net package.[37]

3. Results

Study of clinical predictors such as the number of malaria cases without symptoms and environmental variables, including vegetative index (VI) suggests the seasonal dependences of time series. Appropriate treatment and number asymptomatic variables were included frequently. Temperature and Rainfall were the most common variables in environmental data. The seasonal instances have been analyzed through ACF (Auto correlation) and PACF(Partial auto correlation). Short-term prospects from 1 to 12 weeks have a high frequency of prediction. The RMSPE error was analyzed and found that the geographical location of Khammam has the highest error rate at 117% and Venkatapuram has the lowest error rate as 18.3%. The data series of each geographical location is represented in Table 2.

In this study, we observe that Venkatapuram has the most number of cases. On an average day and night time temperature slightly varies 2 °C from other areas and average rainfall is 1.66 mm with more vegetative index (shown in Table 1). We also observed that Venkatapuram contains number of male malaria cases compared with female cases. The geographical area of Aswaropeta contains a number of predicted cases compared to Khammam and Aswapuram geographical areas. This instance is depicted in Fig. 4. The corresponding input parameters and predicates of all geographical areas are shown in Table 3. Khammam was observed as the smallest number of malaria cases with the average of day and night time temperature of 27.9 °C and rainfall with *1.2 mm* where Aswapuram contains number of female malaria cases compared to the other horizons. All the predicted and observed malaria instances of one year for every two months from *Jan* 2015 *to* 31 *Dec* 2015 are depicted in Fig. 3 as (a), (b), (c), and (d).

4. Discussion

This study observes various variables across four different

Table 2
Error Percentage for prediction of prospects.

Geographical location	Possibility-1 (%)	Possibility-8 (%)	Possibility-28 (%)	Possibility-52 (%)	Average (%)
Venkatapuram	12.2	3.1	26.9	31.0	18.3
Aswapuram	36.5	18.2	58.3	33.0	36.5
Khammam	110.3	127.1	121.0	112	117
Aswaropeta	85.2	65.1	72.6	69.0	72.9

geographical areas in Khammam district and constantly found that both clinical and environmental variables are necessary to attain accurate predictive power. To the best of our knowledge, this is the first study with big data using neural networks other than clinical cases and environmental variables for prediction of malaria abundances. Including clinical variables such as malaria cases with no symptoms confirmed through the lab test that have high predictive variable across the geographical areas. Apart from the clinical variables, environmental variables such as rainfall, temperature, RH, and VI are also identified for achieving maximum results in malaria prediction. The accuracy of the models varies from location to location where the geographical location of Khammam has low malaria cases with high error rate as 117%. There are different processes which can calculate the error and affect the results.[21,22]

Malaria transmission was found to occur between the months of July and September and the peak in the month of August. Rainfall will increase the number of mosquito breeding places which leads to malaria transmission. Here most interesting observation is maximum malaria cases noted between in the month of August and September that shows heavy rains will wash away the mosquito breeding places, as Khammam district start the rains in June.This shows increase in number of malaria cases after July. Relative Humidity and temperature plays important role in malaria transmission when humidity is at 60% and temperature 28 °C makes the mosquito favorable to breed.[2] Where, as when the temperature is greater than 30 °C and lesser than 16 °C may have the negative impact on mosquito breed. From the results, we observe that as venkatapuram,aswapuram,aswaropeta as observed to be a number of cases compared to Khammam because this place contains the high EVI index which supports malaria transmission. Apart from the climatic factors various other factors like type of treatment, level of immunity in human hosts, drug resistance in parasite play important role in malaria transmission. With the malaria prediction, the health facility could plan the patient visits planned for sufficient treatment and diagnoses material available at the time. Concerned authorities who involved in malaria control could understand the impact on endemic places and plan accordingly for coming years. Malaria prediction with big data is the new approach which can lead to dependency on only medical records which can exclude the environmental data in the future. Malaria prediction models are useful in decision-making system where medical facility is limited in practice.

5. Conclusion

In this study, we have provided seasonal patterns of malaria prediction for the geographical sites of Khammam area. We have also confirmed some preliminary findings of malaria abundances by

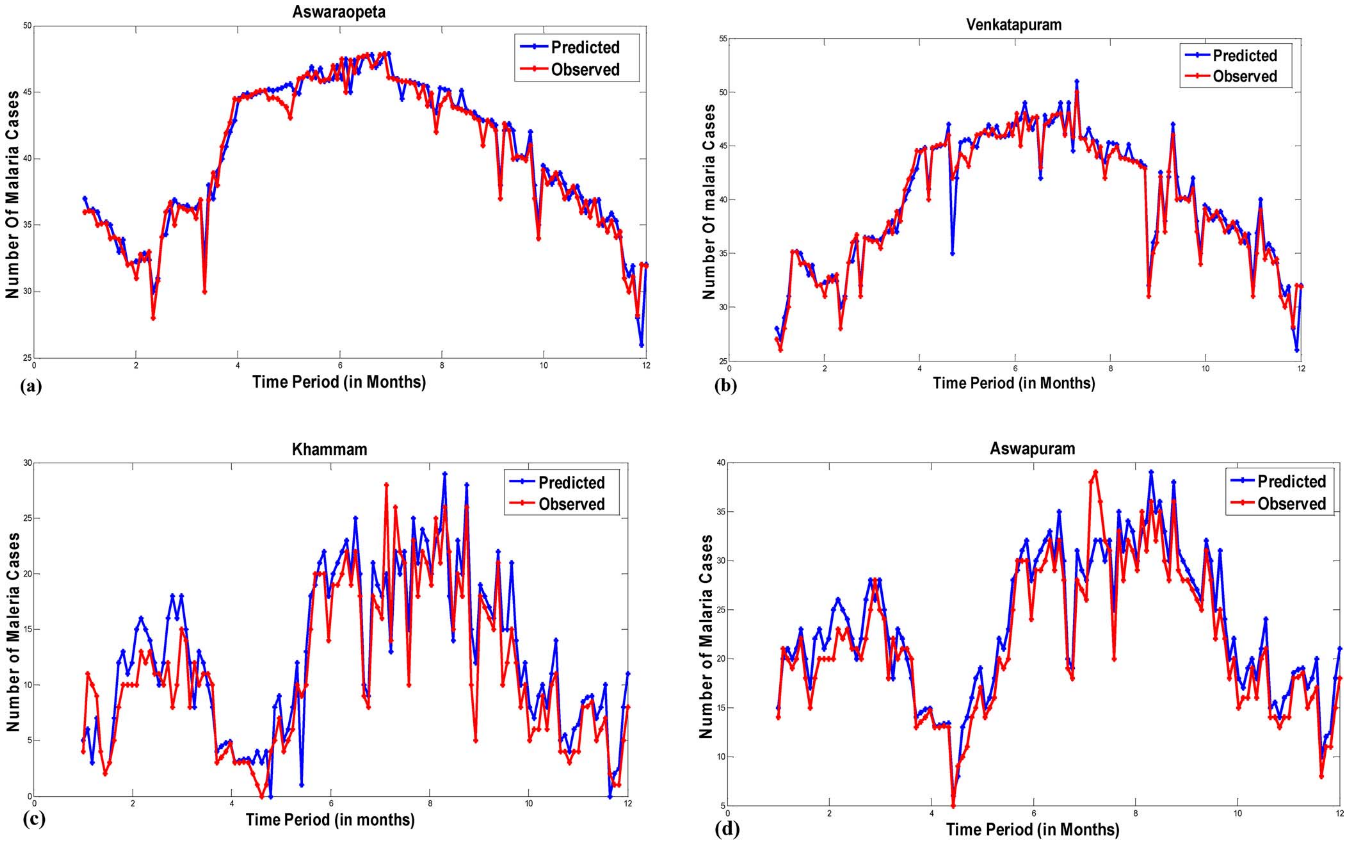


Fig. 3. (a) Predicted and observed malaria instances for ***Aswaraopeta*** Area. (b) Predicted and observed malaria instances for ***Venkatapuram*** Area. (c) Predicted and observed malaria instances for ***Khammam*** Area. (d) Predicted and observed malaria instances for ***Aswapuram*** Area. Predicted and observed instances for all geographical areas.

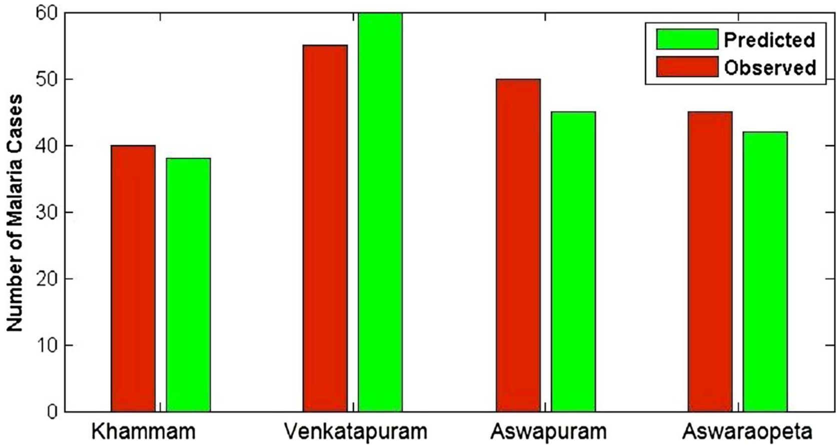


Fig. 4. Predicted vs Observed malaria cases from **Jan 2015 to 31 Dec 2015**.

Table 3
Final input parameters for prediction.

	Venkatapuram	Aswapuram	Khammam	Aswaropeta
Rain fall(RF)	√	√	√	√
Temperature(T)	√		√	√
Relative Humidity (RH)	√	√	√	
Asymptomatic cases (ASC)	√	√	√	√
Vegetative Index(VI)	√	√	√	√

collecting larger data across different seasons. In this study, we have observed climatic conditions and clinical treatment play a significant role in malaria prediction. Till now many models have been developed to examine the malaria prediction with different approaches, but malaria prediction with environmental and clinical data is a new approach to big data analysis. Correct predictive models are required to estimate the disease impact and allocation of medical resources for prevention. More exploration is required in malaria prediction models with better values to bring into real practice.

Acknowledgements

The authors would like to thank Doctor Rambabu, District Medical Officer (DMO), Khammam, Telangana for his support in collecting different seasonal malaria data and Mr. Suresh & Mrs. Mansi, Indian Institute of Remote Sensing (IIRS) for assisting remote sensing data. The authors also like to thank Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, Govt. of India for providing necessary support for carrying the research work.

References

1. Christophers S. Rickard Epidemic malaria of the Punjab: with a note of a method of predicting epidemic years. *Trans Committee Stud Malar India.* 1911;2:17–26.
2. Zinszer Kate, Kigozi Ruth, Charland Katia, Dorsey Grant, Brewer Timothy F, Brownstein John S, Kamya Moses R, Buckeridge David L. Forecasting malaria in a highly endemic country using environmental and clinical predictors. *Malar J.* 2015;14(1):1–9.
3. Gopal Das Nani, Dhiman Sunil, Talukdar Pranab Kumar, Goswami Diganta, Rabha Bipul, Baruah Indra, Veer Vijay. Role of asymptomatic carriers and weather variables in persistent transmission of malaria in an endemic district of Assam, India. *Infection Ecol Epidemiol.* 2015;5.
4. Lauderdale Jonathan M, Caminade Cyril, Heath Andrew E, Jones Anne E, MacLeod David A, Gouda Krushna C, Murty Upadhyayula Suryanarayana, Goswami Prashant, Mutheneni Srinivasa R, Morse Andrew P. Towards seasonal forecasting of malaria in India. *Malar J.* 2014;13(1):1.
5. *Who World Malaria Report 2013.* World Health Organization; 2014.
6. *Forest Report.* 2011; 2011www.forests.ap.gov.in/pdf/APStateofForestReport2011.pdf.
7. Hay Simon I, George Dylan B, Moyes Catherine L, Brownstein John S. Big data opportunities for global infectious disease surveillance. *PLoS Med.* 2013;10(4):e1001413.
8. *Mckinsey.* 2011; 2011www.mckinsey.com/~/media/mckinsey/./mgi_big_data_full_report.ashx.
9. *NCBI.* 2014; 2014http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/.
10. Andreu-Perez Javier, Poon Carmen CY, Merrifield Robert D, Wong Stephen TC, Yang Guang-Zhong. Big data for health. *IEEE J. Biomed. Health. Inf.* 2015;19(4):1193–1208.
11. Hiba Asri, Mousannif Hajar, Moatassime Hassan Al, Noel Thomas. Big data in healthcare: challenges and opportunities. *Cloud Technologies and Applications (CloudTech), 2015 International Conference on IEEE.* 2015:1–7.
12. Barrett Meredith A, Humblet Olivier, Hiatt Robert A, Adler Nancy E. Big data and disease prevention: from quantified self to quantified communities. *Big data.* 2013;1(3):168–175.
13. Boonkiatpong Kritsanatt, Sinthupinyo Sukree. Applying multiple neural networks on large scale data. *Proceedings of International Conference on Information and Electronics Engineering (ICIEE 2011).* 2011.
14. *Neuro Solutions.* 2015; 2015http://www.neurosolutions.com/products/ns/whatisNN.html.
15. Chen An-Sing, Leung Mark T, Daouk Hazem. Application of neural networks to an emerging financial market: forecasting and trading the Taiwan stock index. *Comput Oper Res.* 2003;30(6):901–923.
16. Lee Keun Young, Chung Namil, Hwang Suntae. Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. *Ecol Inf.* 2015.
17. Zhang Guoqiang, Patuwo B Eddy, Hu Michael Y. Forecasting with artificial neural networks: the state of the art. *Int J Forecasting.* 1998;14(1):35–62.
18. Günther Frauke, Fritsch Stefan. neuralnet: training of neural networks. *R Journal.* 2010;2(1):30–38.
19. Birkhead Guthrie S, Klompas Michael, Shah Nirav R. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health.* 2015;36:345–359.
20. Cohen I Glenn, Amarasingham Ruben, Shah Anand, Xie Bin, Lo Bernard. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs.* 2014;33(7):1139–1147.
21. David Van Sickle | LAUNCH. [Online]. Available: http://www.launch.org/innovators/david-van-sickle. [Accessed: 22-dec-2015].
22. Kiang Richard, Adimi Farida, Soika Valerii, Nigro Joseph, Singhasivanon Pratap, Sirichaisinthop Jeeraphat, Leemingsawat Somjai, Apiwathnasorn Chamnarn, Looareesuwan Sornchai. Meteorological, environmental remote sensing and neural network analysis of the epidemiology of malaria transmission in Thailand. *Geospat. Health.* 2006;1(1):71–84.
23. Gao CY, Xiong HY, Yi D, Chai GJ, Yang XW, Liu L. Study on meteorological factors-based neural network model of malaria. *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi.* 2003;24(9):831–834.
24. *Healthcare Data Institute.* 2015; 2015http://healthcaredatainstitute.com/wp-content/uploads/2015/11/hdi_unlocking-the-full-potential-of-data_vf_151125f.pdf.
25. Ram Sudha, Zhang Wenli, Williams Max, Pengetnze Yolande. Predicting asthma-related emergency department visits using big data. *IEEE J Biomed Health Inf.* 2015;19(4):1216–1223.
26. Chai Tianfeng, Draxler Roland R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geosci Model Dev.* 2014;7(3):1247–1250.
27. Revich Boris, Tokarevich Nikolai, Parkinson Alan J. Climate change and zoonotic infections in the Russian Arctic. *Int J Circumpolar Health.* 2012;71.
28. Shone Scott M, Curriero Frank C, Lesser Cyrus R, Glass Gregory E. Characterizing population dynamics of Aedes sollicitans (Diptera: culicidae) using meteorological data. *J Med Entomol.* 2006;43(2):393–402.
29. Ahumada Jorge A, Laoointe Dennis, Samuel Michael D. Modeling the population dynamics of Culex quinquefasciatus (Diptera: culicidae), along an elevational

gradient in Hawaii. *J Med Entomol.* 2004;41(6):1157–1170.

30. Dopazo Joaquín, Wang Huaichun, Carazo José María. *A new type of unsupervised growing neural network for biological sequence classification that adopts the topology of a phylogenetic tree. Biological and Artificial Computation: From Neuroscience to Technology*. Berlin Heidelberg: Springer; 1997:932–941.
31. Sahoo GB, Schladow SG, Reuter JE. Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. *J Hydrol.* 2009;378(3):325–342.
32. Tsai Meng-Hsiun, Yu Shyr-Shen, Chan Yung-Kuan, Jen Chun-Chu. Blood smear image based malaria parasite and infected-erythrocyte detection and segmentation. *J Med Syst.* 2015;39(10):1–14.
33. Shah NK, Dhillon GPS, Dash AP, Arora U, Meshnick SR, Valecha N. Antimalarial drug resistance of *Plasmodium falciparum* in India: changes over time and space. *Lancet Infect. Dis.* 2011;11(1):57–64.
34. Potharaju SP, Sreedevi M. An improved prediction of kidney disease using SMOTE. *Indian J. Sci. Technol.* 2016;9(31).
35. Ramesh D, Suraj P, Saini L. Big data analytics in healthcare: a survey approach. *Microelectronics, Computing and Communications (MicroCom), 2016 International Conference on IEEE.* 2016:1–6 (January).
36. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA.* 2013;309(13):1351–1352.
37. Fritsch S, Guenther F, Guenther MF. *Package 'neuralnet'.* 2016; 2016.