Malaria Outbreak Prediction Model Using Machine Learning

Vijeta Sharma[1],Ajai Kumar[2],Lakshmi Panat[3],Dr. Ganesh Karajkhede[4],Anuradha lele[5]

[1] *Project Engineer,Applied Artificial Intelligence Group ,C-DAC,Pune.*
[2] *Head of Department, Applied Artificial Intelligence Group, C-DAC,Pune.*
[3]*Principle Technical Officer,Applied Artificial Intelligence Group,C-DAC,Pune,*
[4]*Health Informatics Domain Expert,Centre For Development of Advanced Computing,Pune,*
[5] *Joint Director,Applied Artificial Intelligence Group,C-DAC,Pune*

***Abstract*— Malaria is one of the major public health problems in India. Early prediction of a Malaria outbreak is the key for control of malaria morbidity, mortality as well as reducing the risk of transmission of malaria in the community and can help policymakers, health providers, medical officers, ministry of health and other health organizations to better target medical resources to areas of greatest need. Here developed model "Malaria outbreak prediction Model using Machine Learning" can help as an early warning tool to identify potential outbreaks of Malaria. In this study two popular data mining classification algorithms Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used for Malaria prediction using a large dataset of Maharashtra state. Data of all 35 districts of Maharashtra, from 2011 to 2014 has been considered. Parameters used are Average monthly rainfall, Temperature, Humidity, Total number of positive cases, Total number of Plasmodium Falciparum(pF) cases and outbreak occur in binary values Yes or No. A large numbers of samples were collected from different sources. Root Mean Square Error (RMSE) and Receiver Operating Characteristic (ROC) are used to measure the performance of the models. It is observed that performance of the model developed using SVM is more accurate than ANN. The SVM model can predict the outbreak 15 -20 days in advance. However accuracy of prediction can be increased using more training data. This model can be scaled-up at country level.**

***Keywords*— Malaria, Support Vector Machine, Outbreak, Machine Learning, Public Heath,Artificial Neural Network**

I. INTRODUCTION

Malaria is a common disease and sometimes fatal too and that's why it is considered as serious health problem across globe. Malaria is caused by Plasmodium parasites,which are most commonly transmitted through the bite of the Anopheles mosquito. In India, recent studies show that about 95% population in the country resides in malaria endemic areas and 80% of malaria reported in the country is confined to areas consisting 20% of population residing in tribal, hilly, difficult and inaccessible areas [1]. In 1935, 100 million malaria positive patients were diagnosed and 1 milliondeaths occurred. Estimate of 1947 revealed that 75 million cases (21.8% population) occurred in the post-independence population of 334 million with approximately 800,000 deaths [2]. In 1996, India contributed 83% of total malaria cases in South Eastern Region of Asia [3]. All reports indicate that these people could have been saved or treated better if an early warning of this epidemic had been received by health departments of India.

There are several factors which affect the malaria e.g. climate factors (temperature, rainfall, humidity, flood, drought, disasters) [4] and non-climate factors (differences between human hosts, human migration, construction activities). These factors affect the severity of malaria and its transmission. There are many traditional methods used for malaria outbreak prediction e.g. Seasonal forecast model "The Liverpool Malaria Model" - a mathematical-biological model, Auto regressive (AR), Auto-Regressive Moving average(ARMA), Auto-Regressive Integrated Moving average(ARIMA) [5] but accuracy of malaria prediction is always a concern with traditional method and it requires lot of time and effort for data analysis.

Computational model based systems, developed using machine learning techniques are now a days very useful to predict and diagnose many diseases [6-7]. Well defined malaria outbreak parameters [8] are also sufficient to fit in Machine Learning (ML)techniques to effectively and efficiently predict the outbreak [9].As compare to traditional method, these models do not need deep knowledge of statistics. Support Vector Machine (SVM), Naïve Bayes, Decision Tree and Artificial Neural Network (ANN) are some of the major classifiers of Machine Learning techniques which are widely used in healthcare as decision support techniques [10]. Artificial Neural Network work effectively on large number of datasets and input parameters hence it is most commonly used to forecast diseases like cancer[11-12]. Meanwhile Support Vector Machine has proved to be one of the best classifiers for making predictions in two class problem like malaria outbreak(Yes/No)[13].

ISSN: 2278 – 1323

Naïve Bayes is also used as a probabilistic learning method and these classifiers is among the most successful known algorithms for learning to classify text documents like e-mail spam filtering [14]. Some research shows that it is also useful for Heart disease prediction.

Decision Tree machine learning algorithm is also popular for its simplicity and easy touse in decision making and for simple representation, but it requires large training sets to learn and sometime due to lack of enough data it predicts wrong results. Accurate results are highly desirable in health decision making, so this algorithm is not focused in this study [15].

Hence best suited Machine Learning algorithms for health domain - Support Vector Machine and Artificial Neural Network are chosen for building malaria outbreak prediction model and compared for accuracy.

II. METHOD

Below steps shows the method through which malaria outbreak model has developed.

A. Data Collection

Data has been collected from different sources like malaria data from National Vector Borne Disease Control Program, Pune and Meteorological data from Indian Meteorological Department,Pune.Duration of data is from 2011 to 2014,so total 1680 samples were collected for this study. Table 1.1 shows the sample data collected from various sources and Table 1.2 shows testing data to test the model.

Max.Temperature	Min.Temperature	Avg. Humidity	Rainfall	Positive	pf	Outbreak
29	18	49.74	0.00	2156	112	No
34	23	83.27	15.22	10717	677	Yes
40	23	50.74	0.00	1257	127	No
34	24	59.16	9.06	4198	211	No
34	27	73.23	0.00	11808	712	Yes
31	24	88.77	41.40	10881	648	Yes
33	24	77.94	23.88	8830	459	Yes
31	24	84.57	11.15	9693	482	No
36	24	53.40	2.12	9310	549	No
32	23	57.50	0.00	13154	838	Yes
34	18	59.40	0.00	2197	136	No
42	24	49.43	2.19	3362	213	No
45	32	34.74	0.38	416	26	No
43	28	69.07	4.65	7514	410	No
33	23	80.97	6.92	10990	390	Yes
32	24	87.32	11.92	6536	338	No
40	27	63.97	0.00	11169	776	Yes
39	25	47.52	0.00	8131	312	No
36	26	72.78	3.54	5138	213	No
31	23	73.35	4.97	10659	612	Yes
30	23	86.81	7.21	9041	418	No
30	22	78.80	3.12	11265	404	Yes
33	22	73.71	1.75	9233	212	No

Table 1.1: Sample data collected from different sources.

Max.Temperature	Min.Temperature	Avg. Humidity	Rainfall	Positive	pf
35	25	58.14	0.00	5221	110
35	27	65.13	10.09	7452	498
33	24	67.42	4.29	11856	504
32	27	83.19	15.63	10598	614
34	26	81.73	11.34	8432	593
36	26	64.39	4.12	9230	498
35	24	53.87	0.00	10745	453
35	27	84.97	14.55	10639	313

ISSN: 2278 – 1323

34	24	85.48	7.91	11823	549
34	25	81.20	12.08	11276	443
36	24	64.58	0.30	10389	591
32	22	64.30	0.64	8543	365
36	24	69.23	0.00	10545	341
38	22	26.94	0.00	10631	560
38	29	77.43	16.67	11732	462

Table 1.2:Sample testing data prepared to test prediction model

B. Data Preprocessing

District wise malaria data are having different population with respective number of malaria cases. To convert all those raw data into a same format, districts population data have obtain and measure all the districts on same scale. Input variable fields are fixed as each districts average Max temperature, average Min temperature, average rainfall, average mean humidity,number of positive cases, number of pf cases on month wise followed by outbreak reported as output field. There is also provision of handling missing values using "ReplaceMissingValues" feature in Weka (Waikato Environment for Knowledge Analysis) tool. To find a minimal set of attributes that preserve the class distribution, used Weka preprocessor facility to make the parameters on priority consideration using "Select attribute".Finally processed data converted into Weka ARFF(attributerelation file format)to give as input. This malariaTrainingData.arff file become ready to train predictor model .Also,prepared the malariaTestingData.arff for testing the model with missing Outbreak values.

C. Building Model

Weka(Waikato Environment for Knowledge Analysis) data mining tool has been used to simulate data and build predictor model. It is written in java and developed at University of Waikato.Weka is open source, freely available and platform-independent software [17].
Weka has an extensive collection of different machine learning and data mining algorithms and its proven a very helpful data mining tool for developing prediction model by classify the accuracy on the basis of datasets[18].
Weka Explorer has several panel like preprocess, classify, cluster, associate, select attribute and visualize.In this study only "Classify" has used to choose **SVM(Support Vector Machine)** and**ANN(Artificial Neural Network)** classification algorithms. Performance of chosen classifier on dataset shows as WEKA's statistical output in "classifier output" panel.By analyzing below performance parameters,Classifier can be said as best for given dataset-

a.. Correctly Classified Accuracy
It shows the accuracy percentage of test that is correctly classified.

b. Incorrectly Classified Accuracy
It shows the accuracy percentage of test that is incorrectly classified.

c. Mean Absolute Error
It shows the number of errors to analyze algorithm classification accuracy.

d. *Time*
It shows how much time is required to build model in order to predict disease.

e. ROC Area
Receiver Operating Characteristic19 represent test performance guide for classifications accuracy of diagnostic test based on: excellent (0.90-1), good (0.80-0.90), fair (0.70-0.80), poor (0.60-0.70), fail (0.50 – 0.60).

Support Vector Machine(SVM): Support Vector Machine is a supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilisticbinarylinear classifier [19].
Given some training data D, a set of n points of the form

$$D=\{(X_i y_i) \mid X_i \in R^p, Y_i \in \{-1,1\}\}_{i=1 \text{ to } n}$$

Where the Y_i is either 1 or −1, indicating the class to which the point X_i belongs. Each X_i is a P-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_{i=}1$ from those having $y_{i=}-1$. Any hyperplane can be written as the set of points x satisfying,

$$w \cdot x - b = 0,$$

where ' . ' denotes the dot productand w the (not necessarily normalized) normal vectorto the hyperplane. The parameter $b/\|w\|$ determines the offset of the hyperplane from the origin along the normal vector w.
If the training data are linearly separable,we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations
$w \cdot x - b = 1$, and
$w \cdot x - b = -1$

ISSN: 2278 – 1323

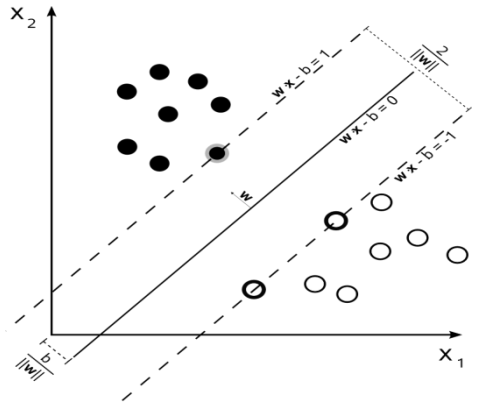

Fig.1 : Support Vector Machine

Artificial Neural Networks(ANN): Artificial Neural Networks is a family of models inspired by biologicalneural network (the central nervous systems of animals, in particular the brain) and are used to estimate or approximatefunctions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning [20] .

An ANN is typically defined by three types of parameters:

A. The interconnection pattern between the different layers of neurons.

B. The learning process for updating the weights of the interconnections.

C. The activation function that converts a neuron's weighted input to its output activation.

Mathematically, a neuron's network function $f(x)$is defined as a composition of other functions$g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables. A widely used type of composition is the nonlinear weighted sum,

$$f(x)=K(\sum_i w_i g_i(x))$$

where, K (commonly referred to as the activation function) is some predefined function, such as the hyperbolic tangent.

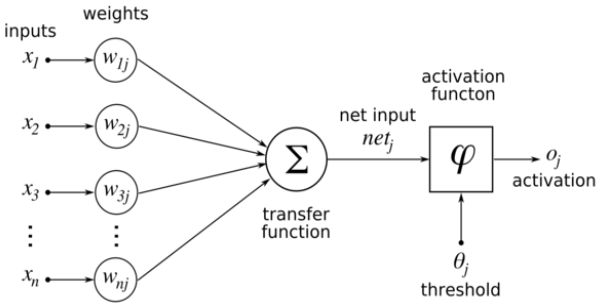

Fig. 2 Artificial Neural Networks(ANN)

In Weka explorer,SVM is listed as LibSVM in classifiers.Firstly,SVM has chosen to build and train the model.So,malariaTrainingData.arffhas supplied aspreprocess file to train the model. Testing file malariaTestingData.arff also supplied using "Supplied test set" in Weka to test accuracy against Output field "Outbreak" in value YES or NO. In libSVM, polynomial kernel type has chosen to achieve the higher accuracy and run the model to observe performance.

ANN shows as "Multilayerperceptron" in classifiers list.Same training dataset malariaTrainingData.arff supplied and used malariaTestingData.arff as test dataset.

D.Choosing best predictor

Root Mean Squared Error (RMSE) and ROC (Receiver Operating Characteristic) performance parameters of SVM and ANN model considered for the comparison of accuracy Model build by SVM gave Root Mean Square Error 0.12 and ROC area 0.89, while ANN produced Root Mean Squared Error 0.47 and ROC area 0.77.As per performance guide for classifications accuracy, it shows that ROC > 0.80 is considered "GOOD" classifier and ROC 0.77 as "FAIR". Classifier should achieve ROC value closer to 1 for higher accuracy of making prediction. For given set of test data SVM model is giving higher number of correct YES or NO value as malaria prediction output against certain values of input parameters as compare to ANN.

Instance No.	observed Value	ANN Prediction	SVM Prediction
1	No	No	No
2	No	No	No
3	Yes	No	Yes
4	Yes	Yes	Yes
5	No	Yes	No
6	No	No	No
7	Yes	No	No
8	Yes	Yes	Yes
9	Yes	Yes	Yes
10	Yes	Yes	Yes
11	No	No	No
12	No	No	No
13	Yes	Yes	Yes
14	Yes	Yes	Yes
15	Yes	No	Yes

Table 1.3: Comparison of ANN & SVM

III. Result And Conclusion

Analysis of comparison table 1.3 shows that performance of Support Vector Machine is more accurate than ANN for the specified testing data set and sample training data set used in this study. This study was based on the idea that values used for various parameters affecting malaria varies as per geological spread.SVM model which is having low error rate has proven to be useful in malaria outbreak prediction among other prediction techniques .

Method	RMSE	ROC
ANN	0.47	0.77
SVM	0.12	0.89

Table 1.4 Error measured between predicted & actual value

ISSN: 2278 – 1323

Prediction capacity of SVM based model in 15-20 days advance can help health organizations to early actions to prevention and cure. It is also observed that learning with more sample data set can improve the accuracy with reducing error rate. Using relative values for the parameters for other demographic area, can be scale upto country level. Presently this early epidemic prediction model can operate at district level health centers of Maharashtra state for getting alarming signals and take preventive action before outbreak occurs.

IV. ACKNOWLEDGMENT

Hearty thanks to Dr. Kanchan Jagtap,Joint Director,National Vector Borne Disease Control Program,Yarwada,Pune to provide me such a real time malaria data,without which my research part could not be completed. I am also grateful to Mr. B.N. Sathe, Officer, National Data Centre, Indian meteorological department,Pune to providing me required meteorological data.

I extend my sincere gratitude to Ms. PallaviGavali,Senior Technical Officer,CDAC,Pune for providing excellent tips and advices, help and encouragement throughout the course of this study.

V. REFERENCES

[1] National Vector Borne Disease ComtrolProgramme,Directorate General of Health Science,Ministry of Health & family walefare,India http://nvbdcp.gov.in/malaria3.html

[2] A Profile of National Institute of Malaria Research, "Estimation of True Malaria Burden in India".

[3] Shiv Lal, G.S. Sonal, P.K. Phukan,"Report-Status of Malaria in India",Journal of Indian Academy of Clinical Medicine Vol. 5 No. 1 .

[4] Paul Edward Parham1 and Edwin Michael, "Modeling the Effects of Weather and Climate Change on Malaria Transmission", Environmental Health Perspectives • volume 118 | number 5 | May 2010

[5] AlemayehuMidekisa,GabrielSenay,Geoffrey M Henebry,PaulosSemuniguse,andMichael C Wimberly,"Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia",MalariaJournal,Volume 11

[6] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar, T Pandu RangaVital,"Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms", International Journal of Engineering and Innovative Technology (IJEIT) ,Volume 3, Issue 3, September 2013.

[7] Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, "EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES".

[8] Rashmi Sharma,"EPIDEMIOLOGICAL INVESTIGATION OF MALARIA OUTBREAK IN VILLAGE SANTEJ, DISTRICT GANDHI NAGAR (GUJARAT)",Indian J. Prev. Soc. Med. Vol. 37 No. 3& 4 , 2006

[9] Orlando P. Zacarias and Henrik Boström,"Predicting the Incidence of Malaria Cases in Mozambique Using RegressionTrees and Forests,",International Journal of Computer Science and Electronics Engineering (IJCSEE) ,Volume 1, Issue 1 (2013)

[10] http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms.

[11] Dr. N. Ganesan ,Dr.K. Venkatesh ,Dr. M. A. Rama,"Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data".

[12] Burke HB,Artificial neural networks for cancer research: outcome prediction,SeminSurg Oncol.1994 Jan-Feb;10(1):73-9.

[13] Orlando P. Zacarias,HenrikBostr¨om,omparing Support Vector Regression and RandomForests for Predicting Malaria Incidence in Mozambique,2013 International Conference on Advances in ICT for Emerging Regions (ICTer): 217 – 221

[14] http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf

[15] Kokol P1,Zorman M,StiglicMM,MalèiaeI,The limitations of decision trees and automatic learning in real world medical decision making.Stud Health Technol Inform. 1998;52 Pt 1:529-33.

[16] https://mahades.maharashtra.gov.in/ppUpdateView.do?report_id=PCA-2013-0005&mode=4

[17] http://www.cs.waikato.ac.nz/ml/weka/

[18] KASHISH ARA SHAKIL, SHADMA ANIS AND MANSAF ALAM , DENGUE DISEASE PREDICTION USING WEKA DATA MINING TOOL

[19] https://en.wikipedia.org/wiki/Support_vector_machine

[20] https://en.wikipedia.org/wiki/Artificial_neural_network

ISSN: 2278 – 1323