Malaria Outbreak Prediction using Machine Learning

Agranee Jha
CS student, dept. Computer Engineering
VCET Mumbai, India

Sanchit Vartak
CS student, dept. Computer Engineering
VCET Mumbai, India

Kavya Nair
CS student, dept. Computer Engineering
VCET Mumbai, India

Prof. Anil Hingmire
Assistant Professor, dept. Computer Engineering
VCET Mumbai, India

***Abstract*—India is prone to a multitude of diseases due to the huge population, area and lack of development. Early prediction of these diseases is the key to controlling the mortality rates and helping in the control of the spread of the disease. Predicting the probabilities of the occurrence of diseases will allow the population to be aware of the risks possible and take preventive measures. Additionally medical resources and aid can be made available to those who require it as early as possible. In this study we will be using machine learning algorithms like Support vector machine to predict the possibility of occurrence of diseases malaria in yes or no class. This study will focus on the possibility of occurrence of diseases with climatic conditions that they have been established to have a relationship with**.

Keywords — Malaria, Support Vector Machine, Outbreak, Machine Learning, Public Heath, Epidemic, Artificial Intelligence, Prediction.

I. INTRODUCTION

Malaria has been a problem in India for quite sometime. Malaria has been around for a long time but during the 1990s it made a comeback with different and new features that were not seen as majorly in the period of the pre-eradication days. These are the vector resistant to different materials such as insecticides, that are called as exophilic vector behaviour, the large amount of vector breeding locations that occurred due to the water resource development projects, growing urbanization and increasing industrialization.[5] Based on the statistics published by World Malaria Report in 2017, in the year 2016, a majority of the population, that is almost around 698 million people were at risk contradicting of malaria. Based on this very Report, India was responsible for 6% of the complete malaria occurrences in the world, 6% of the fatalities, and 51% of the the world over P. vivax cases.[3] The Report estimates the total cases in India at 1.31 million (0.94-

1.83 million) and deaths at 23990 (1600-46500) and that 90 per cent of the deaths were recorded in rural areas, of which almost 86 per cent occurred at home without any kind of medical attention.

According to the National Vector Borne Disease Control Programme (NVBDCP) to assess India‟s actual malaria death burden, the total annual number of cases in India may be about or more than 9.7 million, with about 30,014 – 48,660 deaths (40,297 on an average). There are various factors which are major reasons that cause malaria for e.g. climatic factors like temperature, rainfall, humidity etc and non-climatic factors like different human hosts, human migration etc.[11]

Malaria is caused by Plasmodium Falciparum carried by antelopes mosquito[4]and is spread by the antelopes mosquito Mosquitoes carry diseases that are vector borne and these diseases usually have a relationship to the climatic conditions because of the vector borne nature of the disease.[13]

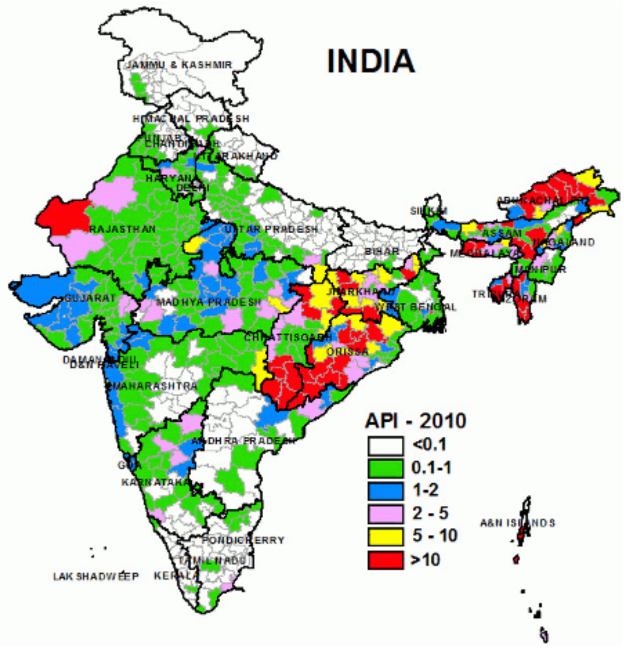


Fig 1[2]

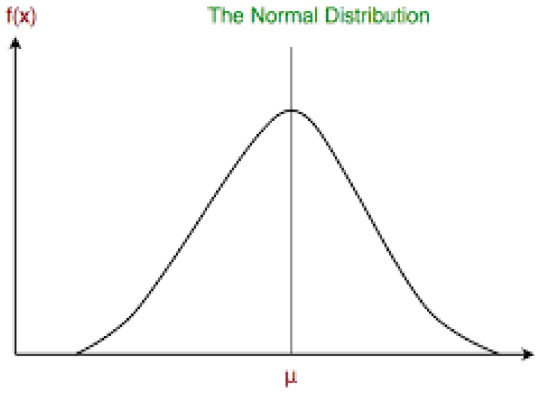
Machine learning and artificial intelligence is a useful prediction methodology. Various techniques are being used to make predictions in different healthcare fields. Clearly given disease outbreak parameters are quite sufficient for the Machine Learning (ML) decision support techniques to correctly give an outbreak prediction. Support Vector Machine(SVM),ARIMA models, Artificial Neural Network are some of the major classifier type of Machine Learning techniques which are widely used in healthcare. Support Vector Machine has in various different situations proved to be one of the most useful classifiers for making predictions in problems with two classes like malaria outbreak that is Yes or No class.

II. POSSIBLE METHODOLOGY

1. Naïve Bayes Method.

Bayes‟ Theorem finds the likelihood of an occasion occurring given the likelihood of another event that has already occurred. Bayes‟ theorem is expressed mathematically because the following equation: Suppose B and A are events and the probability of event B is P(B)=0.

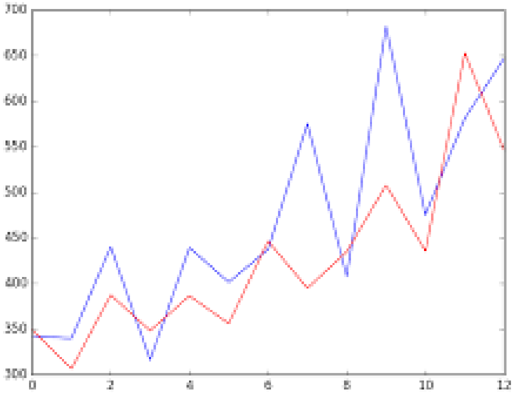
- Basically, we have a tendency to are attempting to seek out probability of an event A, given the event B is true. Event B is additionally termed as proof.
- P(A) is the priori of A (the prior likelihood, i.e. likelihood of event before proof is seen). The proof is associate attribute worth of associate unknown instance(here, it's event B).
- P(A|B) may be a posteriori likelihood of B, i.e. likelihood of event once proof is seen.[9]



2. ARIMA

ARIMA, short for „Auto Regressive Integrated Moving Average‟ is really a category of models that „explains‟ a given statistic supported its own past values, that is, its own lags and therefore the lagged forecast errors, in order that equation is accustomed to forecast future values .Any „non-seasonal‟ statistic that exhibits patterns and isn't a random white noise that is modeled with ARIMA models. An ARIMA model is characterised by three terms: p, d, q where, p is that the order of the AR term, q is that the order of the MA term,d is that the range of differencing needed to form the statistic stationary.[8]

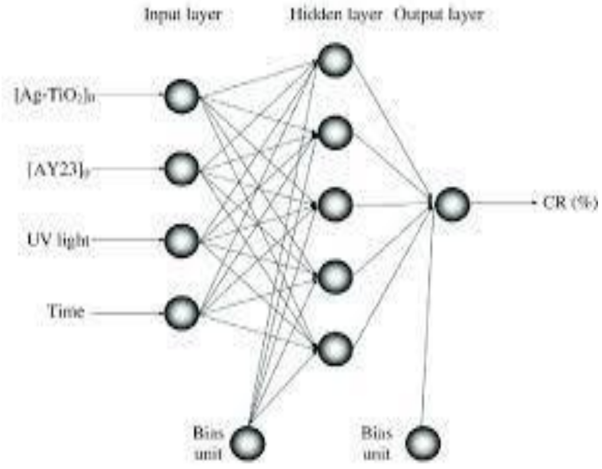
If a statistic, has seasonal patterns, then you would like to feature seasonal terms and it becomes SARIMA, short for „Seasonal ARIMA‟‟. a lot of on it once we have a tendency to end ARIMA.



3. Artificial Neural Network

An ANN is predicated on a gaggle of connected units or nodes mentioned as artificial neurons, that loosely model the neurons during a biological brain. every association, just like the synapses during a biological brain, will transmit an indication to alternative neurons.An Artificial neuron that receives an indication then processes it and may send signal to the neurons connected to that.[10]

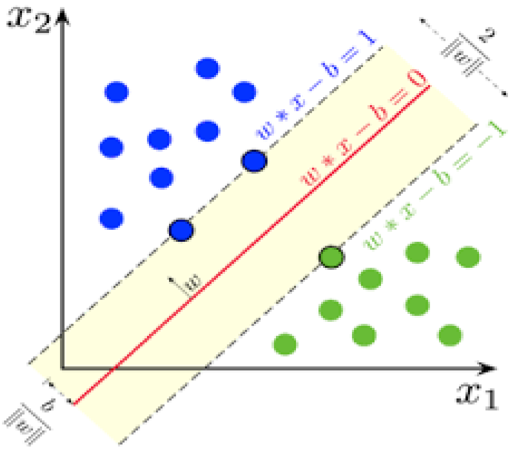
In ANN implementations, the "signal" at the connection may be a real, and therefore the output of every neuron is computed by some non-linear operate of the sum of its inputs. The connections are referred to as edges. Neurons and edges have weights associated with them, these weights help in adjusting the learning rate that is associated with the model. The load will increase or decrease the strength of the signal at an association. Neurons could have a threshold specified an indication is distributed as long as the mixture signal crosses that threshold. Typically, neurons are aggregated into layers. Totally different layers could perform different transformations on their inputs. Signals travel across all the hidden layers present from the input layer and then give an output at the final layer.



II. USED METHODOLOGY Support Vector Machine

Why Support Vector Machine?

Over all the classifiers mentioned above, Support Vector Machine is most commonly used in healthcare and two class situations because of its higher accuracy by establishing a well- defined kernel between two classes and correctly identifying data. A comparison done between ann and svm for a similar malaria based study showed that svm gave a more accurate result.[14]



Support Vector Machine is one of the supervised learning models associated with learning algorithms that analyze information and acknowledge patterns, used for classification and statistical method .If a gaggle of coaching examples square measure given, where every data is marked belonging to one of the 2 classes, associate degree SVM builds a model that assigns new examples into one class or another, creating it a non-probabilistic binary linear classifier.[7]

Given some coaching information D, a group of n points of then

D= Xi ∈ Rp,Yi∈ {-1,1}i=1 to n

Where the Yi is either of the 2 values that's 1 or -1, indicating the category to that the purpose Xi belongs. every Xi as shown may be a P-dimensional real vector. Finding the maximum- margin hyperplane that divides the points having yi=1 from those having yi=-1 being our goal . Any hyperplane is written because the set of points x satisfying the below given condition

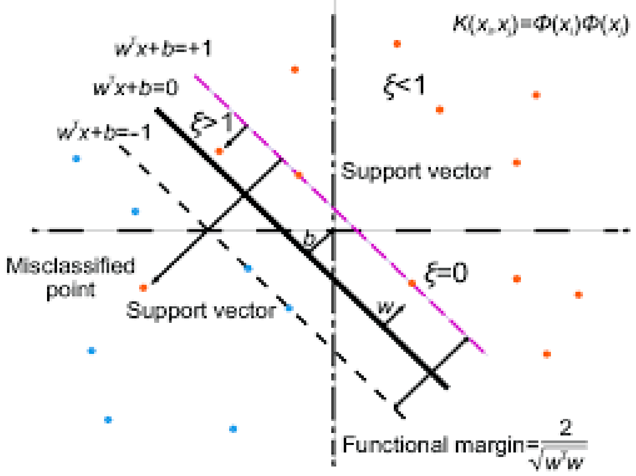
w . x – b=0,

wherein ' . ' denotes the scalar product and „w" the traditional vector to the hyperplane. The parameter b/||w|| determines the offset of the hyperplane from the origin on the traditional vector w.

We can choose 2 hyperplanes if the coaching information square measure linearly divisible ,in a manner that they separate the information and there are not any points between them, and so attempt to maximize the gap between them. The region finite by them is named "the margin". These hyperplanes is more tolerably delineated by the equations given ,

w .x – b=1,

w.x – b=-1.



III. METHOD

Below mentioned steps show the process of developing the malaria outbreak model.

1. Data Collection

   Dataset has been constructed by using sources available at our disposal like malaria data from National Vector Borne Disease Control Program, that gives the number of malaria cases data from 2015 to 2019. Meteorological data from Indian Meteorological Department, the world malaria organization for sample data for different countries. The climatic data has been collected from data.gov.in for climatic conditions in India and WHO for conditions over the world. For the Data not found value “Replace missing Value tool” of Weka has been used.

2. Building Model

   Python is being used to make the prediction model that can be used for purposes in the future. It will be implemented in the jupyter environment to satisfy server requirements. Accuracy comparison is done through implementing the model in weka and obtaining factors.

   Weka(Waikato Environment for Knowledge Analysis) data processing tool is known to act as a simulation for various different models of machine learning . It is has been developed and is being currently used in java. It has been developed at the University of Waikato. Weka is open source, which means that it is freely available and can be used across platforms.[12]

   Weka has an in depth collection of various machine learning and data processing algorithms and its proven a really helpful data processing tool for developing prediction model by

   a.. Correctly Classified Accuracy

   Gives the accuracy percentage of cases that have been correctly classified

b. Incorrectly Classified Accuracy

   It shows the accuracy percentage of test that's incorrectly classified.

c. Mean Absolute Error

   It shows the amount of errors to research algorithm classification accuracy.

d. Time

The amount of proportion time to create this model is given by this factor

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).

Rainfall	Temperature	Humidity	Class
1112	29.17	85.08059336	Yes
1117.6	27.64	59.36520468	Yes
1315.8	28.17	52.47417211	Yes
1124.4	28.12	70.03248155	Yes
2082	27.04	52.27430402	No
3663.9	27.72	87.73328058	No
3443.4	28.11	76.66451781	No
3012.6	25.82	62.81224305	Yes
3461.4	26.81	78.53814755	Yes
4657.2	27.24	27.54365552	Yes
0	22.95	73.04338467	Yes
0	23.9	89.12252669	Yes

IV. RESULT

This paper used Support Vector Machine to try and predict whether malaria outbreak is possible or not based on the climatic factors of temperature, humidity and rainfall being entered.

V. CONCLUSION

Diseases such as malaria usually occur in large numbers and a breakout occurs that can be difficult for all patients and agencies such as hospitals to deal with. Improper handling of such situations by regulatory bodies such as the government leads to mass hysteria among the public and phenomenon such as panic buying. In this paper, we have used SVM to find the whether malaria outbreak can occur in the weather conditions possible or not. For future more data that is further localized can be used to get more accurate predictions. Also, such models can be scaled up to include the country and can be used on different other diseases.

VI. ACKNOWLEDGEMENTS

The author would like to thank Prof. Anil Hingmire for his continued support and guidance. She would also like to thank all his friends and family, without their enthusiasm this project would not have been possible.

VII. REFERENCE

[1] Report on “Estimiation of true malaria burden in India”

[2] https://www.malariasite.com/malaria-india/

[3] https://scied.ucar.edu/longcontent/climate-change-and- vector-borne-disease

[4] World Malaria Health Report, 2017.

[5] Shen, Huajie, Teng Liu and Yueqin Zhang. "Discovery of Learning Path Based on Bayesian Network Association Rule Algorithm." *IJDET* 18.1 (2020): 65-82.
Web. 9 Mar. 2020. doi:10.4018/IJDET.2020010104

[6] Grover-Kopec, E., Kawano, M., Klaver, R.W. *et al.* An online operational rainfall-monitoring resource for epidemic malaria early warning systems in Africa. *Malar J* **4,** 6 (2005).

[7] https://en.wikipedia.org/wiki/Support-vector_machine

[8] https://www.machinelearningplus.com/time-series/arima- model-time-series-forecasting-python/

[9] https://www.geeksforgeeks.org/naive-bayes-classifiers/

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

[11] https://nvbdcp.gov.in/

[12] https://en.wikipedia.org/wiki/Weka_(machine_learning)

[13] 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

[14] Vijeta Sharma, Ajai Kumar, Lakshmi Panat, Dr. Ganesh Karajkhede, Anuradha lele”Malaria Outbreak Prediction Model using Machine Learning”, IJARCET, Volume 4, Issue 12, December 2015.