Malaria Outbreak Prediction Model using Machine Learning

Akash Chandra Patel[1] Anash Shameem[2] Sunil Chaursiya[3] Manish Mishra[4] Abhishek Saxena[5]
[1,2,3,4]B.Tech Student [5]Assistant. Professor
[1,2,3,4,5]Department of Computer Science & Engineering
[1,2,3,4,5]Noida International University, Greater Noida, (U.P.), India

Abstract— There are various diseases whose early detection can prevent the patient from tired and cumbersome medical treatment. Malaria is one of the widely spreading diseases in India [1]. It's early prediction and control is a major challenge for medical science. Various international health community and health based organizations working for its eradication are in greatest need for its prevention. In this paper, we proposed the Malaria Outbreak Prediction Model using Machine Learning which can help us as an outbreak prediction tool to identify potential outbreaks of Malaria. In this study, two popular machine learning classification algorithms Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used for Malaria prediction using a dataset of some cities.

Key words: Malaria, Support Vector Machine, Outbreak, Machine Learning, Public Health, Artificial Neural Network

I. INTRODUCTION

Malaria is the most common disease across tropical regions and sometimes it is very fatal too that's why it is considered a serious health problem all across the globe. Malaria is caused by Plasmodium parasites, which are most commonly transmitted through the bite of the female Anopheles mosquito [2]. In India, 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 the population residing in tribal, hilly, difficult and inaccessible areas [3]. World Health Organization estimates 300–500 million malaria cases annually and in the south-east Asian Region of WHO, out of about 1.4 billion people living in 11 countries (land area 8,466,600 km^2, i.e. 6% of global area), 1.2 billion are exposed to the risk of malaria and most of whom live in India (Kondrachine 1992) [4].

All reports suggest that these people could have been saved or treated better if an early warning of this epidemic had been received by the health departments of India. There are several factors which affect malaria e.g. Temperature, Rainfall, Humidity, flood, drought, disasters [5]. Computational model-based systems, developed using machine learning techniques are nowadays very useful to predict and diagnose many diseases. 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 [6].

A. Support Vector Machine (SVM):

Support Vector Machine is a supervised learning model 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-probabilistic binary linear classifier. 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.x – b = 0, where'. ' denotes the dot product and w the normal vector to 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 [7]

w .x – b = 1, and w.x – b = -1

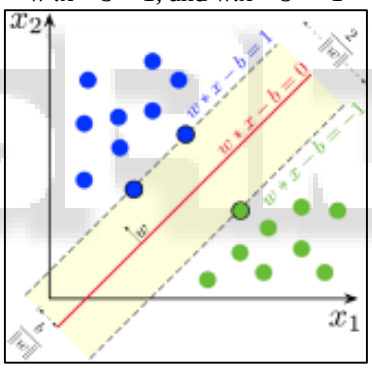

Fig.1: Support Vector Machine

B. Artificial Neural Networks (ANN):

Artificial Neural Networks is a family of models inspired by the biological neural network (the central nervous systems of animals, in particular, the brain) and are used to estimate or approximate functions 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 [8].

An ANN is typically defined by three types of parameters:

1) The interconnection pattern between the different layers of neurons.
2) The learning process for updating the weights of the interconnections.
3) 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.
ANN shows as "Multilayer perceptron" in classifiers list.

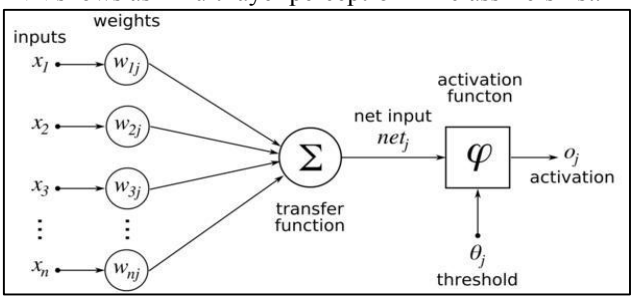

Fig. 2: Artificial Neural Networks (ANN

II. METHODOLOGY

The proposed methodology for the outbreak model has been described as follows:

A. Data Collection

The Data collected for the study is of two years duration 2015-17 from monthly patient data of National Informatics Center and Weather data from various Websites [9]. The total of 120 samples was collected for this study. The following table shows the sample data

	Maximum_Temperature	Minimum_Temperature	Average_Humidity	Rainfall	Total_Case	Pf_Positive	Outbreak
0	27	3	76	16	748	12	Yes
1	31	6	57	11	758	12	Yes
2	40	14	49	11	773	13	Yes
3	44	20	24	4	12	0	No
4	47	21	36	11	40	0	No
5	46	25	50	40	84	0	No
6	39	25	77	234	141	2	No
7	36	24	78	246	263	0	No
8	37	25	67	174	404	0	No
9	36	16	54	46	535	1	No
10	31	12	54	3	646	5	Yes
11	27	7	73	6	683	7	Yes
12	25	7	74	16	1649	22	Yes

Table 1: Sample Data

	Maximum_Temperature	Minimum_Temperature	Average_Humidity	Rainfall	Total_Case	Pf_Positive
0	27	3	76	16	748	12
1	31	6	57	11	758	12
2	40	14	49	11	773	13
3	44	20	24	4	12	0
4	47	21	36	11	40	0
5	46	25	50	40	84	0
6	39	25	77	234	141	2
7	36	24	78	246	263	0
8	37	25	67	174	404	0
9	36	16	54	46	535	1
10	31	12	54	3	646	5
11	27	7	73	6	683	7

Table 2: Actual Testing data to test the model.

B. Data Preprocessing:

The data collected from district wise are having a different population with a respective number of malaria cases. This data has been converted into the same format and are fit to the same scale. Input variable fields are fixed as each district 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 were no missing data so there was no need to handle missing values.

```
Maximum_Temperature      int64
Minimum_Temperature      int64
Average_Humidity         int64
Rainfall                 int64
Total_Case               int64
Pf_Positive              int64
Outbreak                object
dtype: object
```

Fig 3: Data types of variables

B. Building Model

The use of Jupyter Notebook, an interactive tool for data analysis and machine learning has been used to simulate data and build a predictor model. It is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. It is used for data cleaning, transformation, numerical simulation, statistical modeling, data visualization, machine learning.[10]

The outbreak data used in the model is not normalized as optimizing the real observations may affect the accuracy of the model.

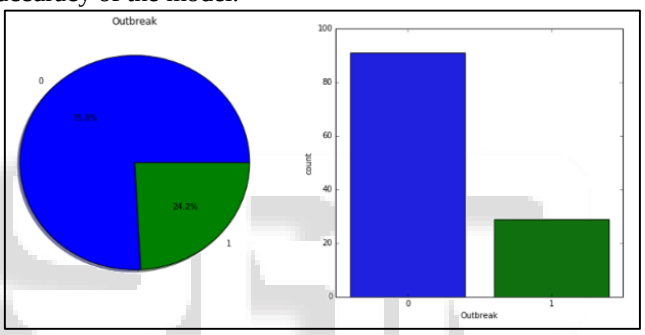

Fig 4: Total Number of positive and negative cases.

C. Choosing Best Predictor

The Similarity score has been used to compare performance parameters of SVM and ANN model considered for the comparison of accuracy. The model build by SVM shows the accuracy of 0.625 while the accuracy predicted by ANN is 0.975. For a given set of test data ANN model is giving a higher number of correct YES or NO value as malaria prediction output, against certain values of input parameters as compared to SVM

III. RESULT

Analysis of a result shows that the performance of Support Vector Machine is less 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 vary as per geological spread. ANN model which is having a low error rate has proven to be useful in malaria outbreak prediction among other prediction techniques.

Method	Accuracy
ANN	0.975
SVM	0.625

Table 3: Accuracy Of models

It also has been observed that Temperature and Humidity play a major role in malaria morbidity.

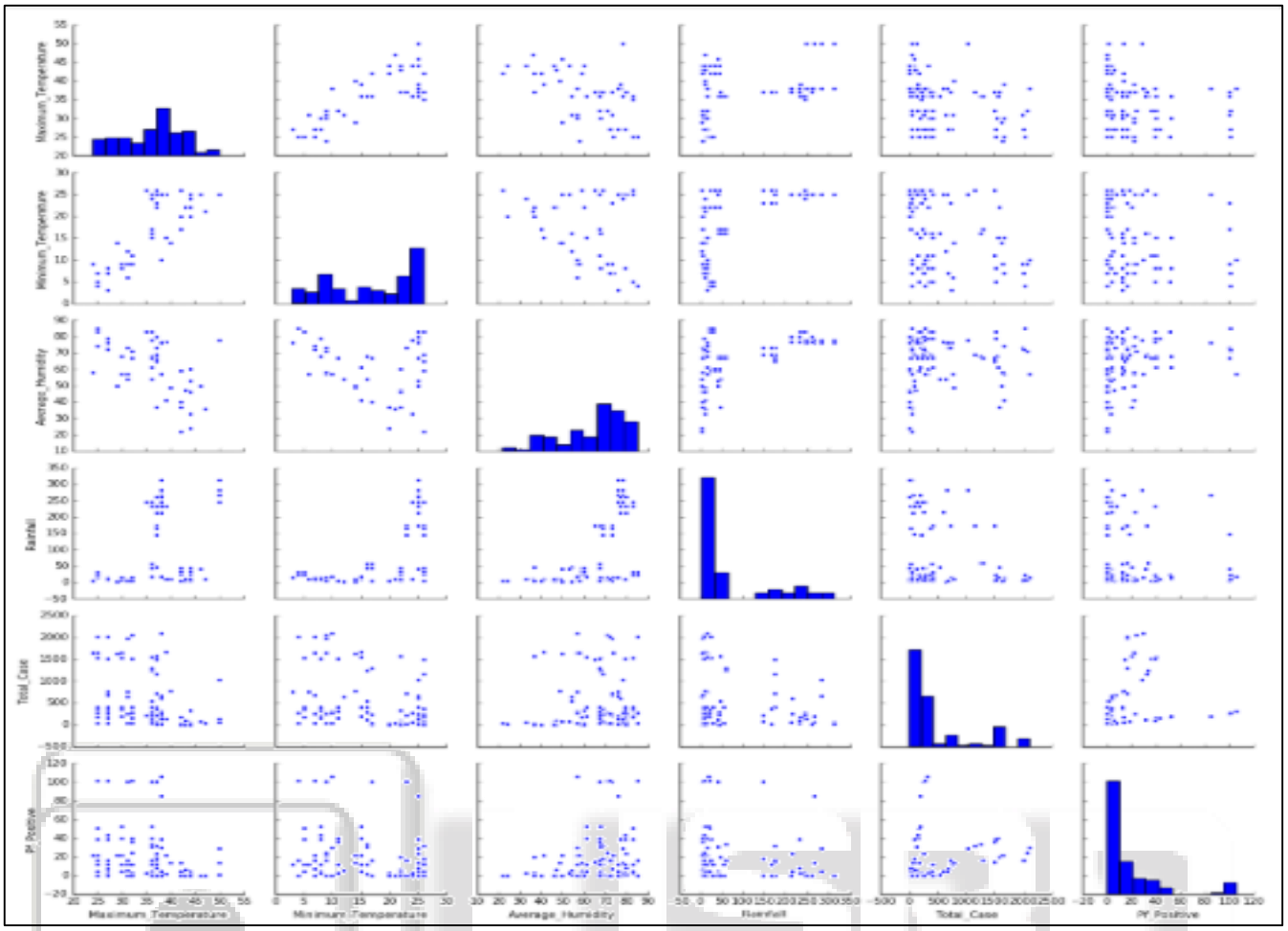

Fig. 5: Features Relation graph

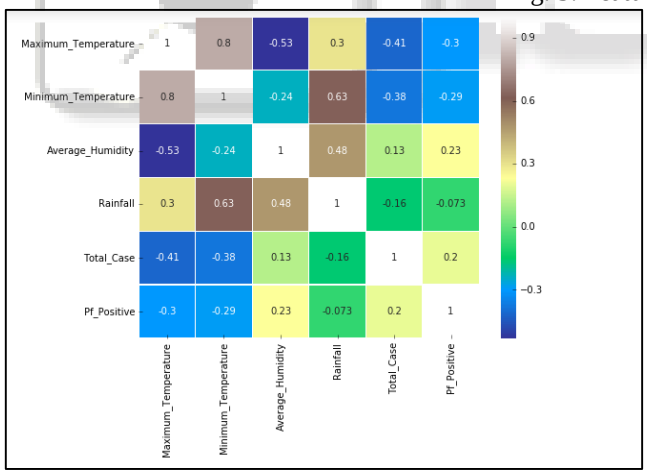

Fig. 6: Heatmap

IV. CONCLUSION

Prediction capacity of ANN-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 the error rate. Using relative values for the parameters for other demographic areas can be scaled up to country level. Presently this early epidemic prediction model can operate at district level health centers of Uttar Pradesh state for getting alarming signals and take preventive action before the outbreak occurs.

REFERENCES

[1] https://www.malariasite.com/malaria-india/
[2] https://www.who.int/features/qa/10/en/
[3] https://www.nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=420&lid=3699
[4] https://www.ncbi.nlm.nih.gov/books/NBK1720/
[5] http://www.open.edu/openlearncreate/mod/oucontent/view.php?id=89&printable=1
[6] https://www.tandfonline.com/doi/full/10.1080/10106049.2018.1489422v
[7] https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72
[8] https://en.wikipedia.org/wiki/Artificial_neural_network
[9] https://www.nic.in/
[10] https://jupyter.org/

