International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Prediction of Epidemic Outbreaks in Specific Region using Machine Learning

A. K. Bhavana, Chalumuru Suresh, B. V. Kiranmayee, Kadem Shravan Kumar

Abstract: ***Dengue is one amongst the common infectious disease which is caused by dengue virus and transmitted to humans by mosquitoes with this many are infected in varied regions around the world per year. The reason for this virus is atmospheric conditions, which plays a vital role in the outbreak of dengue. Therefore early prediction of dengue is the key to regulate outbreak and reduces the transmission within the community. To overcome this we are using various machine learning (ML) algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest tree (RF) and Decision Tree (DT) are used to predict the dengue outbreak. Prediction is done based on weather parameters like monthly wise maximum temperature, minimum temperature, average temperature, mean temperature, humidity and Precipitation which is considered as weather dataset and this weather dataset is pre- processed using label encoding function before applying into the training models. The performances of all the models are calculated based on weather dataset. After considering performance of all the models we choose random forest as a best predictor for dengue outbreak.***

Keywords: ANN, Decision Tree, Machine Learning, Random Forest tree, SVM and Weather parameters

I. INTRODUCTION

An epidemic is the rapid spread of infectious disease within a short period of time, usually two weeks or less, to a large number of people in a given population. There are different types of epidemic diseases outbreak such as dengue, malaria, hepatitis, chikungunya and many more.

Dengue is a most common disease caused by mosquitoes which affecting humans and it is the fastest emerging viral disease in many parts of the world in recent years. Sometimes it leads to death and thus it is considered as most dangerous disease across the world.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
AK.Bhavana*,Department of Computer Science and Engineering,VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad, India. Email: kannojubhavana@gmail.com
Chalumuru Suresh, Department of Computer Science and Engineering,VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad, India. Email: suresh_ch@vnrvjiet.in
B.V.Kiranmayee, Department of Computer Science and Engineering,VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad, India. Email: kiranmayee_bv@vnrvjiet.in
Kadem Shravan Kumar, Department of Computer Science and Engineering,VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad, India. Email: shravankadem101@gmail.com

This disease is caused by one of the four different types of virus, i.e., DENV1-4, this viruses unfold by Aedes mosquitoes that grow close to human lodgings.

This unfold occurs when a mosquito bites an individual infected with an infectious disease virus then the virus from the infected person enters into mosquito. Once the infected mosquito bites an alternative person then the virus enters into that person. It is causing 50 to 100 million cases yearly. The number of cases of dengue is increasing year by year. There are several factors which cause dengue, they are climate factors like rainfall, humidity, temperature, precipitation, floods, drought, and non-climate factors like human population, human migration, water collected in the containers etc. These factors cause the severity of dengue and its transmission so, to control the transmission we need to predict the outbreak of dengue early. There are several old methods used for prediction of dengue outbreak. For better prediction of outbreak we used machine learning technology.

Machine learning brought advancement in healthcare field which reduces human efforts like manual calculations and avoids mistakes. we use machine learning to predict dengue outbreak for improving precision, and to help the health control agency to enhance its strategic disease control planning to tackle dengue outbreaks in that region.

II. LITERATURE SURVEY

YuhanisYusof*et.al*[1], They developed a model to predict infectious disease incidences for electoral regions at intervals in Sri Lanka. This study used mobile network information to induce the human population quality of Sri Lanka and used totally different information sources like census information and surveys for typical information. Artificial neural networks and XG Boost algorithms applied on this dataset to predict infectious disease. This work concluded that human quality is the main reason for the unfold of break bone fever in Sri Lanka.

AsrulNizam*et.al*[2], Developed an online primarily based system that provides real world applied math analysis and prediction on dengue outbreak natural event for the community.

MarceloAparecidoCarlos*et.al*[3], In this work they used data, information originated from social networks, text mining techniques and machine learning algorithms like k-means and SVM for implementing a model to predict dengue outbreak natural event.

NaiyarIqbal*et.al*[4], They predicted dengue outbreak using various machine learning classifiers ranging from simple classifiers, like Decision Tree, Naïve Bayes, Model Tree, to complex algorithms such as Support Vector Machines,

Neural Networks, Gene Expression Programming, Genetic Programming and ensemble classifiers.

YuhanisYusof*et.al*[5], They have collected dengue fever cases and downfall level from 5 districts in Selangor and this information is preprocessed using point standardization and applied LS-SVM algorithm to predict dengue outbreak.

ShermonS*et.al*[6], Considered Dengue incidence cases and climate factors as dataset, machine learning techniques like k-means clustering and multiple regression used to build a model for prediction.

Guohun Zhu *et.al*[7], In their work dengue cases and rainfall DPE used as a dataset and SVM algorithms used for building model to detect dengue outbreak. This study says that there is a strong correlation between dengue cases and rainfall DPE features.

Mathulamuthu*et.al*[8], Manifold learning theorem used to reduce the dimension, used weather, climate and dengue cases as dataset, machine learning algorithms like k-means and regression to build a model for dengue forecast.

P. Manivanna*et.al*[9], D win's method used for preprocessing the dengue dataset and for attribute selection. Data mining technique like k-means clustering algorithm used for implementing dengue fever predictor.

Azuraliza Abu Bakar*et.al*[10], Here they developed a prognostic model to predict break bone fever eruption mistreatment Multiple Rule based mostly Classifiers. Totally different rule based mostly classifiers used are the choice Tree, Rough Set Classifier, Naive mathematician, and Associative Classifier to make classification modeling for prediction.

ZurianiMustaffa*et.al*[11], Here predicts future dengue outbreak using three different normalization techniques named min-max, z-score, decimal point normalization. These techniques used in prediction model built using LS-SVM and neural network algorithms. This study says that LS_SVM is a better predictor than the neural network model.

Yan WU*et.al*[12], Here the wavelet transformation is used for data pre-processing and for selecting main features Support Vector Machines (SVM) -based Genetic Algorithm is used. To build a model for forecasting dengue, they used SVM regression. This study says that the link between climatic factors and dengue cases are powerful control measure towards dengue outbreak.

S. Chadsuthiet.al[13], They proposed a model for prediction of dengue outbreak using climatic dataset and pre-processing techniques like normalization, data transformation, standardization, feature selection and data cleansing ,algorithms used are decision tree, regression analysis, ANN, SVM, KNN.

M. H. Sulaimanet.al[14], They developed a model for prediction of dengue outbreaks by considering dengue cases, rainfall, temperature as dataset and machine learning algorithm like Least Squares Support Vector Machines.

III. PROPOSED METHOD

Early identification of dengue outbreak is very helpful to control it. So here in proposed method identified dengue outbreak based on the key factors of climatic conditions like max temp, min temp, mean temp, hum, prep of specific regions and developed a method for classification and prediction of dengue outbreak using this key factors data. In our work various ML algorithms like Random forest tree, Decision tree, Support Vector machine, and artificial neural network is used for weather data. Initially, data are pre-processed using different methods for weather data. Next processed data are used for training classification algorithms and for classifying data into 'yes' or 'no' and at the last performance of the algorithms is calculated and compared to know the best predictor algorithm. Proposed method contains different steps they are data collection, Data pre-processing, Building models for prediction and choose best predictor. Below diagram represents the flow of project:

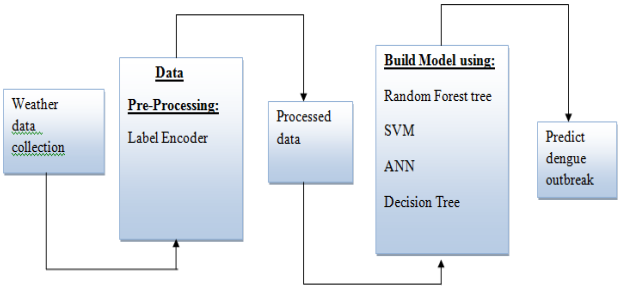

Fig 1: Prediction of dengue outbreak

Step 1: Data collection

Dengue outbreak data and weather data has been collected from data driven website for a period of 18 years and those datasets are mapped with month attribute. Total 419 samples were collected for this study. Below table represent the sample data.

Table I: Climatic conditions with outbreak dataset

year	dewpoint_temp	precipitation(mm)	max_air_temp_c	Humidity	month	Outbreak
1991	292.74	3.9625	299.475	74.91	1	No
1991	293.08	4.5125	300.025	76.55	2	No
1991	293.45	6.358	300.32	78.26	3	No
1991	293.07	24.0525	299.275	76.77	4	No
1991	294.88	22.675	300.6	80.06	5	No
1991	295.64	37.51	301.78	79.67	6	No
1991	296.03	29.77	302.2	79.82	7	No
1991	295.93	29.3275	301.9	79.79	8	yes
1991	295.28	48.565	301.7	79.77	9	yes
1991	293.86	29.3275	300.12	77.87	10	No
1992	293.981	67.99	300.0333	81.23	1	Yes

Step 2: Data Pre-processing

Data preprocessing is an associate degree integral step because the quality data directly affects the flexibility of our model and also helps is achieving higher accurate results.

Therefore, it's important to preprocess our data before feeding it into our model. In this work Preprocessing is done using preprocessing function

Label Encoder: It is used to convert categorical information or text information, into numbers, that our prognosticative models will higher perceive.

Data set contains all numeric values except target attribute i.e., outbreak so for better results we used label encoder to convert "yes" and "no" to '1' and '0'In Below table we can see the sample for conversion.

Table II: Dataset after pre-processing

year	dewpoint_temp	precipitation (mm)	max_air_temp_c	Humidity	month	Outbreak
1991	292.74	3.9625	299.475	74.91	1	0
1991	293.08	4.5125	300.025	76.55	2	0
1991	293.45	6.358	300.32	78.26	3	0
1991	293.07	24.0525	299.275	76.77	4	0
1991	294.88	22.675	300.6	80.06	5	0
1991	295.64	37.51	301.78	79.67	6	0
1991	296.03	29.77	302.2	79.82	7	0
1991	295.93	29.3275	301.9	79.79	8	1
1991	295.28	48.565	301.7	79.77	9	1
1991	293.86	29.3275	300.12	77.87	10	0
1992	293.981	67.99	300.0333	81.23	1	1

To split the pre-processed data into training and testing we imported train_test_split() function from sklearn.model_selectionand alienated data into 80% training set and 20% test set. The training set is used to train the model. Testing set is used to test the trained model. Output field outbreak in numerical value 1 for "YES" and 0 for "NO". Performance is calculated using metrics Precision, Recall and F1-score.

Step 3: Building model using various algorithms

1) Random forest Random forest is a type of supervised ML algorithm based on Ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times i.e., multiple decision trees to form a more powerful prediction model. In Random forest, each tree in the ensemble is built by replacing training set sample.In our work we implemented random forest classifier by importing Random Forest Classifier() from sklearn. ensembleand this classifier is used with two parameters, n_ estimator and Random_ State and it is trained using training climatic dataset. n_ estimator defines number of trees we taken n_ estimator as '20'and random state is random number generator we taken Random_ sate as'0' for training.

```
RandomForestClassifier(n_estimators
='20',Random_state='0')
```

After defining classifier we fitted the classifier by using fit function with trained input data and trained output data. Finally Trained classifier is used to classify outbreak into 'yes' or 'no' using test data set.

2) Support Vector Machine: SVM is a supervised ML algorithm which can be used for classification or regression problems. We used for classification. It uses a technique called the kernel trick to transform your data and then based on these transformations, it finds an optimal boundary between the possible outputs. SVM is a classifier which can segregates two classes with the help of hyper plane. In our work we implemented support vector classifier by importing SVC from sklearn.svm and used linear kernel as shown in below equation.

```
svclassifier = SVC(kernel='linear')
```

SVM classifier is trained using training data, this trained SVM classifier is used to predict whether the given climatic conditions show outbreak or not.

3) Artificial neural network: ANN is generally presented as systems of interconnected "neurons" which exchange messages between each other. The connection has numerical weights that can be
tuned based on experience, making neural nets adaptive to input and capable of learning. In our work we used a multilayer perceptron classifier (MLP) which is imported from sklearn.neural_networkand it is used with hidden layers size as 13 and maximum iterations as 500.

```
MLPClassifier(hidden_layer_sizes=(1
3,13,13),max_iter=500)
```

This classifier fit the model using fit function with trained input and out data and predicts output. Here we send new data to predict the outbreak as 'yes' or 'no'.

4) Decision Tree: Decision tree is one of the predictive modelling algorithm used to split the data based on the given conditions. This process is recursive in nature and is repeated for every sub tree rooted at the new nodes. It is used for both classification and regression here we used decision tree for classification.From sklearn.tree we imported decision tree classifier to classify the climatic dataset as outbreak or not. In this algorithm we used two attributes, they are max_depth and random state. The level of depth is 3 and random state as 1.

```
DecisionTreeClassifier(max_depth=3,
random_state=1)
```

Based on this attributes we trained the dataset and then we fit the model. Later this model predicts the outbreak as yes or no on the new dataset.

Step 4: Choose best predictor:

Performance parameters of SVM, ANN, RF, DT models are considered in the comparison of accuracy. Model build with good accuracy is chooses as a best predictor.

IV. RESULTS

One of the most critical and final steps in building a model is to assess model efficiency. In this study, precision, Recall, accuracy score, F1-score, support are considered to be performance indicators for choosing best model for prediction. For performance evaluation, we took 20% of total data which is also known as a test data .These values are predicted using learned model.

For each model, the accuracy score is calculated by taking in to account the actual result and the predicted value. As state d below, the mathematical method for the calculation of the accuracy score.

$$Accuracy = TP+TN/TOTAL$$

Precision is the percentage that corresponds to the measurem ents of actual results represent a proportion of the true positi ve. The mathematical method for the calculation is as shown below.

$$Precision=TP/TP+FP$$

Recall is the ratio of true positives and total positives. The mathematical method for the calculation is as shown below.

$$Recall=TP/TP+FN$$

The harmonic mean of recall and precision is F1-score. As s hown below, the mathematical method for calculating F1-sc ore is

$$F1\text{-}score=2*precision*recall/precision+recall$$

We consider 335 samples of climatic conditions dataset and trained Random forest, decision tree, ANN, SVM with this dataset and for testing data we consider 84 samples.

Fig[2] demonstrate actual values of outbreak for particular climatic conditions, here x-axis is number of observations and y-axis is outbreak

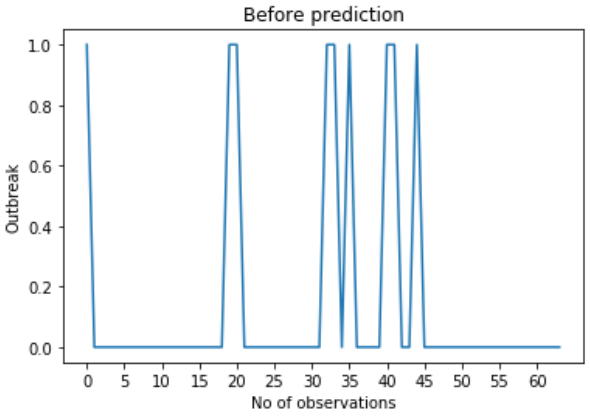

Fig 2: Outbreak of actual values

As we used four different algorithms, we analysed prediction of outbreak using this four algorithms fig[3] shows prediction of dengue outbreak using SVM ,fig[4] shows prediction of dengue outbreak using ANN, fig[5]shows prediction of dengue outbreak using Decision Tree, fig[6] shows prediction of outbreak using Random Forest. '0' in figures represents there is no outbreak for particular Climatic conditions and '1' represents for outbreak for particular climatic condition

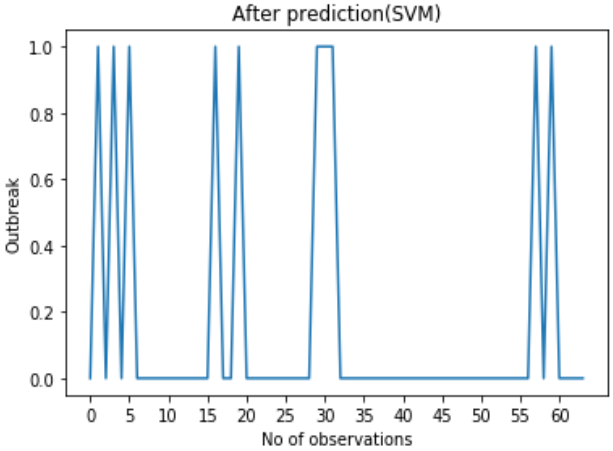

Fig 3: Predicted outbreak using SVM

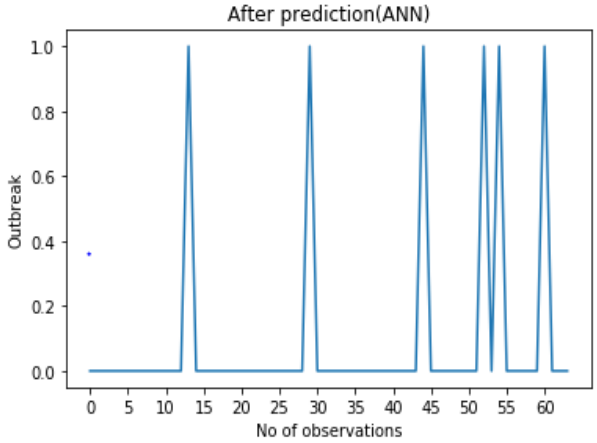

Fig 4: Predicted outbreak using ANN

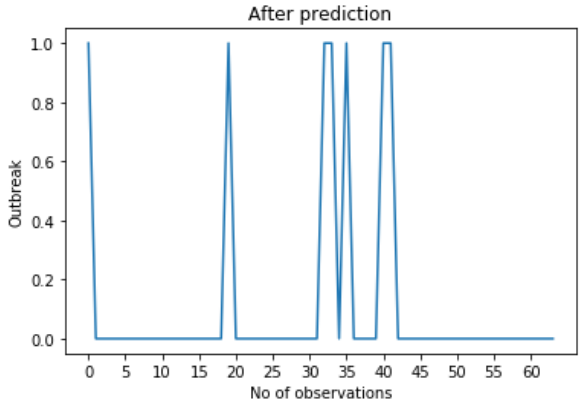

Fig 5: Predicted outbreak using DT

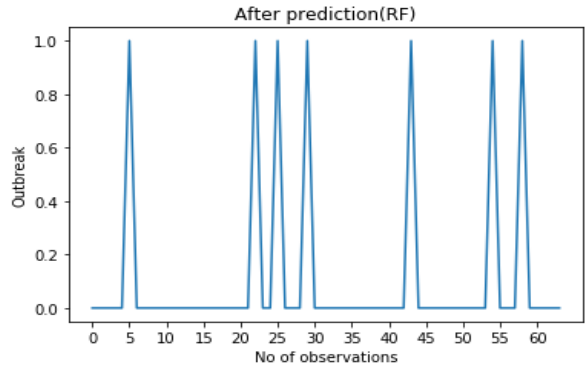

Fig 6: predicted outbreak using RT

Table III represents performance metrics Precision, Recall, F1-Score, Support and accuracy rate for both normal condition and outbreak condition for all algorithms for SVM. Similarly Table IV demonstrate performance metrics for ANN. Performance metrics of Decision Tree shown in Table V, Table VI represents for Random Forest.

Table III: Performance of SVM

	Precision	Recall	F1-score	Support
0	0.81	1.00	0.90	52
1	0.00	0.00	0.00	12
WeightedAvg	0.66	0.81	0.73	64
Accuracy	0.8125			

Table IV: Performance of ANN

	Precision	Recall	F1-score	Support
0	0.96	0.85	0.90	52
1	0.56	0.83	0.67	12
Weighted Avg	0.88	0.84	0.85	64
Accuracy	0.84375			

Table V: Performance of DT

	Precision	Recall	F1-score	Support
0	0.89	1.00	0.94	57
1	0.00	0.00	0.00	7
Weighted Avg	0.79	0.89	0.84	64
Accuracy	0.890625			

Table VI: Performance of RF

	Precision	Recall	F1-score	Support
0	1.00	0.98	0.99	58
1	0.86	1.00	0.92	6
Weighted Avg	0.99	0.98	0.98	64
Accuracy	0.984375			

Fig[7] Demonstrate comparison of all four classification algorithm that is Support vector machine, Random forest, Decision tree, Artificial neural network, In below graph it shows that Random forest has highest accuracy of 98% and accuracy decision tree got 89% which is after Random forest , Artificial neural network got accuracy 84% and Support vector machine got accuracy 81% which is low compare to in four algorithms.

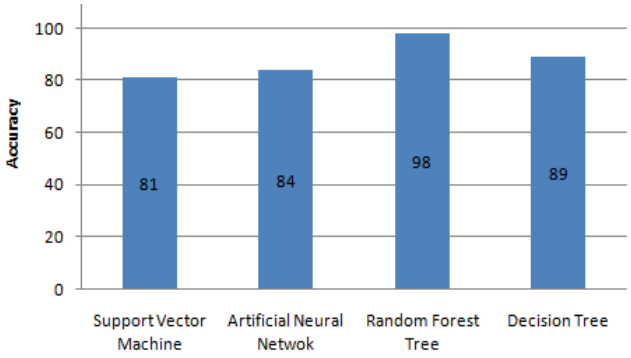

Fig 7: Comparison of algorithm

Table VII: Accuracy comparison table

Algorithms	Accuracy percentage
SVM	81%
ANN	84%
Random Forest	98%
Decision Tree	89%

V. CONCLUSION

In our work we focused on prediction of dengue outbreak based on weather parameters like monthly wise maximum temperature, minimum temperature, average temperature, mean temperature, humidity and Precipitation, after our analysis we found that above factors are effecting the outbreak of dengue, we processed these parameters using label encoder and implemented machine learning algorithms like Support vector machine, Random forest, Decision tree, and Artificial neural network individually to predict outbreak of dengue disease, The performance of all these algorithms are calculated using accuracy and observed that SVM, ANN, RF, DT got 81%, 84%, 98%, 89% respectively. Compared to all the algorithms, random forest is giving the best accuracy than other models.

Random Forest, which is a higher accuracy rate proven to be useful in dengue outbreak prediction among other prediction models, In future we can implement the random forest model with IOT devices in which climate sensors can read the above climatic conditions and predict the outbreak of dengue.

REFERENCES

1. K. G. S. Dharmawardana et al., "Predictive model for the dengue incidences in Sri Lanka using mobile network big data," 2017 IEEE International Conference on Industrial and Information Systems (ICIIS), Peradeniya, 2017, pp. 1-6.
2. A. N. Asat, A. F. Mahat, R. Hassan and A. S. Ahmed, "Development of dengue detection and prevention system (Deng-E) based upon open data in Malaysia," 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), Langkawi, 2017, pp. 1-6
3. M. A. Carlos, M. Nogueira and R. J. Machado, "Analysis of dengue outbreaks using big data analytics and social networks," 2017 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, 2017, pp. 1592-1597. 2017.
4. Naiyar,MohammadIslam,"Machine learning for dengue outbreak prediction: An outlook",IJARCS,2017
5. YusofandZuriani,"Dengue Outbreak Prediction: A Least Squares Support Vector Machines Approach",IJCTE,2016
6. S. S. Mathulamuthu, V. S. Asirvadam, S. C. Dass, B. S. Gill and T. Loshini, "Predicting dengue incidences using cluster based regression on climate data," 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 245-250.
7. G. Zhu, J. Hunter and Y. Jiang, "Improved Prediction of Dengue Outbreak Using the Delay Permutation Entropy," 2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Chengdu, 2016, pp. 828-832.
8. S. S. Mathulamuthu, V. S. Asirvadam, S. C. Dass and B. S. Gill, "Predicting dengue cases by aggregation of climate variable using manifold learning," 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, 2017, pp. 535-540.
9. P. Manivannan and P. I. Devi, "Dengue fever prediction using K-means clustering algorithm," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, 2017, pp. 1-5.
10. A. A. Bakar, Z. Kefli, S. Abdullah and M. Sahani, "Predictive models for dengue outbreak using multiple rulebase classifiers," Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, Bandung, 2011, pp. 1-6.
11. Zuriani, yuhanisyusof"A comparison of normalization techniques in predicting dengue outbreak",IACSIT,2011
12. Yan WU, Member, IAENG, Gary LEE,"Detect Climatic Factors Contributing to Dengue Outbreak based on Wavelet, Support Vector Machines and Genetic Algorithm",ISBN: :978-988-98671-9-5, 2016
13. P. Siriyasatien, S. Chadsuthi, K. Jampachaisri and K. Kesorn, "Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes," in IEEE, 2018
14. Z. Mustaffa, M. H. Sulaiman, F. Emawan, Y. Yusof and M. F. M. Mohsin, "Dengue Outbreak Prediction: Hybrid Meta-heuristic Model," IEEE, 2018
15. H. Somwanshi and P. Ganjewar, "Real-Time Dengue Prediction Using Naive Bayes Predicator in the IoT", IEEE, 2018.

AUTHORS PROFILE

A. K. Bhavana, Currently pursuing Master of Technology in Computer Science and Engineering in VNR Vignana Jyothi Institute of Engineering and Technology affiliated to JNTU Hyderabad. Completed Bachelor of Technology in department of Computer Science and Engineering from JNTU. Her works related to Machine learning.

Chalumuru Suresh, currently working as Assistant Professor in the department of CSE at VNR Vignana Jyothi Institute of Engineering and Technology affiliated to JNTU Hyderabad. He is having good experience in teaching and research. His research areas include Data Mining, Big Data Analytics and Machine Learning.

B. V. Kiranmayee, Associate Professor in Department of Computer Science and Engineering, VNRVJIET, Hyderabad .she is currently pursuing Ph.D. degree in Computer Science and Engineering in JNTUH University.

Kadem Shravan Kumar, Currently pursuing Master of Technology in Computer Science and Engineering in VNR Vignana Jyothi Institute of Engineering and Technology affiliated to JNTU Hyderabad. Completed Bachelor of Technology in department of Computer Science and Engineering from JNTU .His works related to Machine learning.