RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 04, Issue 04 April
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2022.025

# A Data Mining based study on Dengue Fever: A Review

*Kousik Bhattacharya*[1], *Dr. Avijit Kumar Chaudhuri*[2], *Dr. Anirban Das*[3], *Dilip K. Banerjee*[4]

[1]*Research Scholar, Department of Computer Application, Seacom Skills University, West Bengal, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Techno Engineering College Banipur, Kolkata, West Bengal, India*
[3]*Professor, University of Engineering and Management, Kolkata, West Bengal, India*
[4]*University Research Professor, Seacom Skills University, West Bengal, India*

Emails: kousik8086@gmail.com, c.avijitresearch@gmail.com, anirban-das@live.com, dkbanrg@gmail.com

## Article History

*Received: 21 March 2022*
*Accepted: 24 April 2022*

***Keywords***:

*Dengue;*
*Malaria;*
*Chikungunia;*
*Typhoid;*
*COVID19*

## Abstract

*Dengue fever (DF) is amosquitoborne disease spread by female Aedes mosquito. Dengue transmission depends on the changing of climatic parameters like temperature, humidity, rainfall, as well as the congestion in an area, i.e., where the population density is high. In this review, we have highlighted the reasons of the occurrence of DF and methods for early detection of the same. Symptoms are the key points to diagnose the dengue patients. Many diseases like Malaria, Chikungunia, Typhoid, COVID-19, etc. have the common symptoms of fever, body pain, eye pain, diarrhoea, etc. Few rare symptoms have been identified for diagnosing DF using machine learning predictive model. Rare symptoms are skin disease, headache, abdominal pain for early detection of dengue.*

## 1. Introduction

Dengue is a viral fever caused by a bite of female Ades mosquito. Dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS) are the two deadly forms of dengue disease. Most of the cases of dengue are either DSS or DHF has been continuously reported all over India. Four different types of DHF are seen among the dengue patients, which are DHFI, DHFII, DHFIII, and DHFN. Most of the researchers have used Data Mining (DM) techniques in their studies. DM is also useful technique to analyze the different factors like health care services, environmental, and agricultural, and food etc. DM is an essential technique-based applications to discover for diagnosing of DF. The machine learning algorithms like Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) classification have used for predictive analysis on DF.

## 2. Review of Literature

In the research work of Mello-Roman, Gomez Guerrero, Torres, 2019, dataset was made on the basis of collection of data from the admitted patients of Paraguay from 2012- 2016, for early diagnosis of dengue disease. Two machine learning techniques, Artificial Neural Network (ANN) and Support Vector Machine (SVM) were used to compare medical early diagnosis. Tests were completed with the help of the IBM SPSS Modular Software, where classification models were used on the 90% training dataset and 10% test dataset. Using SVM technique, with an average of 90% accuracy, Sensitivity, specificity, but in comparison ANN polynomial obtained better results of 96% accuracy, 96% sensitivity, 97% specificity in thirty random partitions of the dataset with

low variations (Mello-Román et al.).

Symptoms like joint pains, rash, metallic taste, vomiting, the headache, of dengue affected people are be shown within three to fourteen days. Deaths of the dengue patients are caused more due to a lack of early diagnosis of disease. Researchers used a dataset consist of 58 number of disease ailments, and they were designed a model using Bayes Server (BS), a machine learning technique to detect the dengue haemorrhagic fever (DHF); by showing various kinds of symptoms which are related to other diseases of fever. After testing the data of the DHF medical repository using BS had 99.84% preaccuracy.

In the study of Fathima, Manimeglai, 2012, the authors, using data mining computational analytic and SVM (support vector machines) used to identify the abroviral disease – dengue and to identify patterns and rules for making decisions about future work, implementing the data sets, the best accuracy is very satisfactory (Fathima and Manimegalai). But it is very time-consuming because of more parameters. It needs more calculation time. Data are integrated from multiple sources. But data and processes are dynamic. The flexibility of design will make sensed data inviolable for retrospective analysis.

In a previous study, the authors, using the Decision Tree Approach, consisting of a set of selected dengue attributes, qualified by the gained ratio and Data Mining Method to correctly classified the dengue patients and to detect the day of defervescence of fever which is called day 0, the critical data of dengue patients who face the fatal condition, reached the result that decision tree approach did not suit to obtain accuracy; and wanted to select a new classification approach. In the paper, a decision tree, one of the vital data mining tools was used. To classify dengue patients, the decision tree approach gave the researchers good results, but to focus day 0 prediction, and it gave low accuracy.

In another study, authors conducted an exploration of the Dengue outbreak in Pondicherry. They aimed to detect the members and details of people affected by fever during the outbreak period and also to get the environmental factors. They used a community- based cross- sectional investigative study using pretested questionnaire Data regarding the age, sex, education, occupation, economic status, history of fever, laboratory investigation, hospitalization, verification of available reports on diagnosis and treatments, mosquitoes breeding places from each patient. It was discovered that discarded tires, coconut shells, flower vases, uncovered barrels and buckets, household water storage vessels were containing mosquito larvae in the affected place. Eight incidents of dengue fever were informed from the semi town area of Pondicherry, but no cases of death were reported. Daily waged people were most affected in comparison with affected people of other occupations.

The various Data Mining techniques in health care such as classification, clustering, association, and regression are analyzed. The necessity of efficient analytical methodology for tracking out new and essential information in health- related data for the health industry, health insurance to check fraud, availability of medical solution to the patients at a lower cost, detection of cases of diseases, identify effective medical treatment methods, and efficient health care policies. By data mining technique, it is adequate to analyze factors responsible for diseases such as food, various working environments, the educational level along with the living condition, available ability of freshwater, health care services, cultural, environmental, and agricultural factors. The authors warned against giving guidelines for using the data mining technique. No single classifier can produce the best result for every data set. This data set consists of training and testing. The performance of a classifier is judged using the testing data set. But sometimes, a testing data set may be easy and sometimes complicated. To avoid this problem, cross validation may give good performance in both training and testing. A hierarchical clustering technique is used where there is less information. Besides Dendrograms, partitioned algorithm is analyzed for overcoming the shortcomings of clustering. Association is useful for identifying relationships among various attributes. An insignificant association is removed by experts.

Authors have previously studied to find the environmental conditions conducive for the outbreak of Dengue fever, to trace the spatial variations of the disease in different parts of Kolkata, to identify the socio-economic grounds behind the amplitude of the disease in slum areas of Kolkata, to know about the variations of the outbreak of the disease among

people based on their housing conditions, to assess the role of government and NGO; in regulating the wideness of the illness, to understand the level of awareness among ordinary people about the danger of mosquito bite which leads to dengue fever. In doing so, the researcher has assumed the adjacent areas of Kolkata especially parts of Howrah, North & South 24 Parganas, three administrative decisions of West Bengal comparing the incidents of Dengue fever with the rest of India in respect of spread and cases of fatality rate. According to the authors, a more careful approach must be taken to combat the disease, community participation is required in urban and rural areas, and awareness building has to be more intensive to check the mosquito generation.

Dengue virus is reportedly a virus of the group Flavivirus of the species Flavivviridae, which includes four types of dengue fever- DEN1, DEN2, DEN3, and DEN4. According to them, dengue is a sickness of tropical and subtropics countries. The dengue upsurge is for the development of population growth rate, unplanned urbanization, inadequate mosquito control, numerous air travel, and scarcity of health awareness facilities. Dengue gets into more than 100 countries including, Europe &USA. The authors are of the view that in 1780 the first virologically certified epidemic dengue come in Calcutta and Eastern India between1963-1964. Dengue fever is a flue like an infectious disease, attacks persons of all ages, and it occurs chiefly during the rainy season. It is spread by Aedes mosquito bite. Dengue virus infection gives identified clinical response. So, its accurate diagnosis is very difficult before clinical test. Antivirus of dengue is not discovered; physicians of the prescribed analgesic medicine as supportive care, fluid intake and sufficient bed rest.

Researchers have categorized DHF, on the basis of different symptoms; the first category is Dengue Fever (DF). Symptoms of DF are same as Typhoid Fever (TF). The symptoms of second category DHF are fever, nausea, vomiting, red spots, and nose bleeds. The third category is dengue shock syndrome (DSS). It is the final or level-3 stage of DF. In this level, it can be affected in heart, brain, lungs, kidney, and also in that case patients feels breathing, and fainting problem. Classification study was conducted to identify the stage of DHF disease, and helps to doctor to diagnose. ID3 classic algorithm for the dataset including decision tree is used in respect of symptoms of the level of DHF, and achieved accuracy of 82% (Rosid et al.).

The author (Niriella et al.), analyzed on the case record of 697 number of dengue patients, for early detection of clinical phase (CP) of dengue which is helpful for doctor to diagnose the patients. Here Logistic Regression was implemented to identify independent risk factor for CP. 226 number of patients were fall in CP out of 697 numbers of patients in the unit. $\chi^2$ and t-test were used to compare with the categorical and continuous variables respectively. From analysis, it was concluded, that positive independent predictor of CP (OR 2.83) and negative predicted value: 97.2%.

The author analyzed (Arafiyah and Hermin), established, that to avoid misdiagnosis by the Doctor, right prediction, for DF treatment using ANFIS has an application programme, which is dedicated, a patient has DF or not. With the use of NB, the studies predicted that on the basis of inputted clinical data of temperature, spotting, bleeding, and rumple. And the output variables containing suffers from DBD, or diagnosis of the patients who are suffering from DHF or not. Result of the model test, achievement level of precision is 77.3%.

Authors studied (I Nordin et al.), collected data of dengue cases from Health Department of Ketantan, Malaysis for predict of dengue. They established a prediction model built on the basis of three Kernel functions along with Gaussian radial basis function (RBF), by using SVM for predicting future design outbreak. Result obtained the highest prediction accuracy of 85% .

Authors have also studied the environmental and socio-economic risk factor of dengue fever. Spatial analysis, including point density, average nearest neighbor, Spatial autocorrelation, hot spot analysis, were used to analyze and Spearman rank correlation, Ordinary least Square (OLS), were used to investigate the environmental, Socio-Economic risk factors of dengue fever. They experimented on 30553 cases of dengue fever of five districts of China in 2014. After case study, it show strong seasonal variations, and most of the cases (96%) of the total areas of dengue patients were found August to October of the year. Most of the cases of the total were found in the high density area, which were located in the districts junctions. The DF was strongly co-related with LT,

normalized difference water index (NDWI), Land Surface temperature of day time (LSTD), Land Surface temperature of night time (LSTN), population density (PD), gross domestic product (GP), where correlation of 0.483, 0.456, 0.612, 0.699, 0.705, 0.205 respectively. Reset of the adjusted R-squared was 0.320 (Yue et al.).

The scholars analyzed (Cheong, Leitão, and Lakes), have shown the use of land depending upon the water bodies or agricultural practices which were the key factors to influence the complex interactions among vector host and virus for transmission of dengue disease. They used Boosted Regression Tree (BST) for predicting highest accuracy. Using this model, result of Cross-Validated performance score (Area under the Receiver Operator Characteristic Curve or ROC_AUC) is 0.81.

In another study, for pre-identification of DF; four machine learning models of pls, glmset, RF, xgboost, were evaluated with testing data set ROC_AUC as the quantitative measure is 0.94 and predicted accuracy was 88% (Salami).

In a recent study (Sahak), tested 569 samples from the month of May to December, 2019, of DENV, where 213 (37.4%) cases positive and 356 (62.6%) cases negative. Symptoms of all the cases were, fever, headache, myalgia, and arthralgia. Clinical features were low plate late (50%), eye pain (36%), rash (21%), and nausea or vomiting (21%). Overall, they used simple mean, and median statistical method to describe epidemiogical characteristic of DENV.

In comparison to others according to the various symptoms of DF and used machine learning algorithm, (Caicedo-Torres, Paternina, and Pinzón), recorded an accuracy level of 95% as well as both sensitivity and specificity of 65% and a ROC_AUC score of 0.75.

## 3. Methodology

### *3.1. Support Vector Machine (SVM)*

SVM is one of the most supervised machine learning models, which can be used in both linear and nonlinear problems. SVMs are very useful method to work on the unknown data which may be unstructured or semi unstructured data like text, images, tress etc. This method is applicable for finding the Optimal Separator function which can be separated dataset into two categories. SVM is also the most linear marginal classifier method which has been used previously for both classification and regression cases.

### *3.2. Decision Tree (DT)*

DT model is a supervised classification tree which has leaves, and branches represent class levels and conjunctions feathers. A decision and decision making both are used in the decision tree. Two types of the decision trees are used in data mining. One is the Classification Tree, and another is the Regression Tree. The testing was conducted on the data, using the DT; method ID3 algorithm, in terms of symptoms that affect DHE; and achieved the highest accuracy.

### *3.3. Logistics Regression (LR)*

LR is a statistical method has two possibilities may be true or false. LR model has two categories, one is multinomial logistics has more than two outputs, and other is ordinal LR. In the LR; method, logistic function which is cumulative distribution function of logistic distributors is worked to measure the probability to maintain the relationship between categorical dependent variable or one, and more than one independent variable. Author used LR is analyzed to detect symptoms, physical signs that classified the DF, and getting 74% sensitivity and 79% specificity.

### *3.4. Naive Bayes (NB)*

NB model is created on the basis of Bayes theorem, In NB model, conditional probability is used, and it embarks posterior class probability for each instance in the data set, by using Bayes theorem.Torres, Paternina, Pinzon, 2016, used Gaussian Priors from NB model were implemented to find each feature mean and estimated Variance. Arafiyah, Hermin, 2018, have taken input data of fever, processing of bleeding, spotting, tourniquet test; they used the NB; model to predict whether or not affected dengue. The performance of the classification NB; algorithm, using ROC; the prediction accuracy is 69% (Chadwick et al.)

### *3.5. Random Forest (RF)*

Random decision forest for Classification and Regression is investigated machine learning algorithm was first raised by Ho in 1995. The first conceptual paper was made on Random Forest by Leo Breiman in 2001. The most popular complex classification technique where supervised of more clas-

sifiers can only increase to certain levels of accuracy, and alleviates errors. Trees increase low bias to very high Variance; both in RF; are of common multiple dense decision trees on various parts of the dataset with reducing Variance. The simple bootstrap aggregating methods can be used for RF; because without increasing the bias, it decreases the inconsistency of the model. RF; are non parametric, and it can handle categorical, and multi-model data which are maybe ordinal or non-ordinal (Chang). In his journal, Fathima, Manimegbi, 2015, has taken 500 number of trees, and 5 number of variables per split is used for the RF; Classification method to measure the importance of predictor variables, accuracy; in mean decrease and, Gini index. Arafiyah, Hermin, 2018, proved, based on the result system, from their data set of patient's medical records, the algorithm RF; with classification accuracy is much better than where they have used measure the performance difference and got AUC_RF accuracy rate is in between 0.80 – 0.90. Based on the result of accurate DHF; prediction system for avoiding the error of diagnosing DHF.

### 3.6. AdaBoost

AdaBoost stands for "Adaptive Boost". After selecting the training subset for accurate prediction of the last training the algorithm repetitively trains the AdaBoost model and it will be continuing for the strong probability of classification from the second in order of repetition or iteration, it gives higher gravity to wrong classified supervision. This process will continue unless and until training data assemble without any error.

### 3.7. Cohen's Kappa (CK)

CK is a statistical procedure to measure of the reliability of two raters give the same rating. The reliability of raters depends on the number of agreement scores. According to Kappa statics, CK, K has measured the agreement between categorical variables x and y.

If the value is

1. 0 agreement to chance
2. 0.10 – 0.20 slight agreement
3. 0.21 – 0.40 fair agreement
4. 0.41 – 0.60 moderate agreement
5. 0.61– 0.80 substantial agreement
6. 0.81 – 0.99 near perfect
7. 1 perfect

To calculate K, authors have used SPSS software. Formulation for Cohen's Kappa,

K= $P_0$ – Pe / (1-$P_0$), where probability of agreement $P_0$ = (Number in agreement / Total)

And, Pe = A(correct) + A(incorrect)

$P\ (correct) = (A\ +\ B\ /\ A\ +\ B\ +\ C\ +D)\ *\ (A\ +C\ /\ A\ +B\ +\ C\ +\ D)$.....(1)

$P\ (correct) = (C\ +\ D\ /A\ +\ B\ +\ C\ +D)\ *\ (B\ +\ D\ /\ A\ +\ B\ +\ C\ +\ D)$ .....(2)

Where, A is the total number of raters is correct. The raters are in agreement.

B is the total number of rater 1 is incorrect, but rater 2 said are correct, this has disagreement.

C is the total number of rater 2 is incorrect, but rater 1 said are correct, this has disagreement.

D is the total number of both raters are incorrect; this is agreement.

### 3.8. ROC_AUC

Full form of ROC is Receiver Operating Curve, and AUC is Compute Area Under. There is comparing with two operating characteristics, True Positive Rate (TPR), and False Positive Rate (FPR), in ROC; is also called the relative operating characteristic curve. ROC; curve has structured by marking the TPR is also called sensitivity; or in machine learning, it is known as the probability of detection, against the FPR; which is treated as probability of false alarm, and it has computed as (1-Specificity). For ROC, the AUC must be used roc_auc_score()function. Both the roc_curve and the AUC function take true outcomes, i.e. (0, 1), and enumerated the class 1.

LR-> Logistic Regression, SVM-> Support Vector

Machine, LSTM-> Long Short Term Memory, MSO-> Multi Swam Optimization, MLP-> Multilayer Perception, ANN-> Artificial Neural Network, NB-> Naïve Bayesian, DT-> Decision Tree, RF-> Random Forest, BBN-> Bayesian Belief Network.

## 4. Conclusion

From the study of different review papers, and made the table-1 and it has observed that, result of the accuracy label, above 90% is more affective, only using specific machine learning models, of ANN & SVM, or BBN, or RF. From the analysis of data of various review papers, we found, specific reason for affecting people in Dengue, what are the symptoms

**TABLE 1. Comparative Paper Study with data analysis**

| In Paper | Accuracy | $R^2$ | Sensitivity | Specificity | Precision | $F^2$ | ROC_AUC | TPR & FR | Used Model |
|---|---|---|---|---|---|---|---|---|---|
| 1. | 96% | - | 96% | 97% | | | | | ANN & SVM |
| 2. | 92.19% | | 94.04% | 92.19% | | | | 0.51 & 0.99 | RF |
| 3. | 82% | | | | | | | | DT |
| 4. | | | | | 77.3% | | | | NB |
| 5. | 85% | | | | | | | | SVM |
| 6. | | 0.320 | | | | | | | Spearman rank Correlation |
| 7. | | | | | | | 0.81 | | Boosted Regression Tree |
| 8. | 88% | | | | | | | | RF |
| 9. | 85% | | 74% | 86% | | | 0.90 | | SVM |

of dengue affected people, and what damage occurs in the human body. The authors conclude that, a seasonal prevailing wind in the region of South and South East Asia blowing from South West brining rain from the month of May to September. This season is advantageous for fertilization of the Aedes mosquito, which is spreading the dengue virus, and so these countries of South and South East Asia are mostly affected in dengue. Symptoms of the dengue patients for detection of DF are body pain, vomiting, head ace, cough, loose stool, not sufficient to detect dengue; these, factors are now the similar symptoms to other diseases like Malaria, Chikungunia, Typhoid, and COVID-19. The author include, more symptoms of eye pain or red eye or both, hiccups, are added as major reason for early detection of DF. And very important conclusion in the paper, the damage may be occurred in the Liver, Prostate, Spleen, and different Spot shown on the human body, Enzyme system failure, the Nerve system failure in the brain and dengue patients suffers from blood sugar, weight loss, appetite and weakness after recovering from DF.

## ORCID iDs

Kousik Bhattacharya https://orcid.org/0000-0003-1810-0563

## References

Arafiyah, Ria and Fariani Hermin. "Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method". *Journal of Physics: Conference Series* 948.1 (2018): 012077–012077. 10.1088/1742-6596/948/1/012077.

Caicedo-Torres, William, Ángel Paternina, and Hernando Pinzón. "Machine learning models for early dengue severity prediction". *Ibero-American Conference on Artificial Intelligence* (2016). 10.1007/978-3-319-47955-2_21.

Chadwick, David, et al. "Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: Application of logistic regression analysis". *Journal of Clinical Virology* 35.2 (2006): 147–153. 10.1016/j.jcv.2005.06.002.

Chang, Ko. "Dengue fever scoring system: new strategy for the early detection of acute dengue virus infection in Taiwan". *Journal of the Formosan Medical Association* 108 (2009): 879–885. 10.1016/S0929-6646(09)60420-4.

Cheong, Yoon Ling, Pedro J. Leitão, and Tobia Lakes. "Assessment of land use factors associated with dengue cases in Malaysia using Boosted Regression Trees". *Spatial and Spatio-temporal Epidemiology* 10 (2014): 75–84. 10.1016/j.sste.2014.05.002.

Fathima, A and D Manimegalai. "Predictive analysis for the arbovirus-dengue using svm classification". *International Journal of Engineering and Technology* 2.3 (2012): 521–528. 10.1.1.411.9082.

I Nordin, N, et al. "The Classification Performance using Support Vector Machine for Endemic Dengue Cases". *Journal of Physics: Conference Series* 1496.1 (2020): 012006–012006. 10.1088/1742-6596/1496/1/012006.

Mello-Román, Jorge D., et al. "Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay". *Computational and Mathematical Methods in Medicine* 2019 (2019): 1–7. 10.1155/2019/7307803.

Niriella, Madunil A., et al. "Identification of dengue patients with high risk of severe disease, using early clinical and laboratory features, in a resource-limited setting". *Archives of Virology* 165.9 (2020): 2029–2035. 10.1007/s00705-020-04720-5.

Rosid, M A, et al. "Classification Of Dengue Hemorrhagic Disease Using Decision Tree With Id3 Algorithm". *Journal of Physics: Conference Series* 1381.1 (2019): 012039–012039. 10.1088/1742-6596/1381/1/012039.

Sahak, Mohammad Nadir. "Dengue fever as an emerging disease in Afghanistan: Epidemiology of the first reported cases". *International Journal of Infectious Diseases* 99 (2020): 23–27. 10.1016/j.ijid.2020.07.033.

Salami, Donald. "Predicting dengue importation into Europe, using machine learning and model-agnostic methods". *Scientific Reports* 10 (2020): 1–13. 10.1038/s41598-020-66650-1.

Yue, Yujuan, et al. "Spatial analysis of dengue fever and exploration of its environmental and socio-economic risk factors using ordinary least squares: A case study in five districts of Guangzhou City, China, 2014". *International Journal of Infectious Diseases* 75 (2018): 39–48. 10.1016/j.ijid.2018.07.023.



**Embargo period:** The article has no embargo period.

**To cite this Article:** Bhattacharya, Kousik, Dr. Avijit Kumar Chaudhuri, Dr. Anirban Das, and Dilip K. Banerjee. "**A Data Mining based study on Dengue Fever: A Review** ." International Research Journal on Advanced Science Hub 04.04 April (2022): 101–107. http://dx.doi.org/10.47392/irjash.2022.025