RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 04, Issue 05 May
www.rspsciencehub.com

http://dx.doi.org/10.47392/irjash.2022.034

Comparison of Recent Data Mining Algorithms to Identify of the factors and effects of Dengue Fever and Ensemble Random Forest, A new Algorithm

Kousik Bhattacharya[1], *Dr. Avijit Kumar Chaudhuri*[2], *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: 3 March 2022
Accepted: 21 May 2022

***Keywords*:**

Dengue;
AdaBoost;
Ensemble;
Random Forest;
SVM;
RUC_AUC

Abstract

The disease dengue has created panic in the minds of men and women of this time. Now a day the menacing of dengue has spread from town areas to rural areas. It affects heavily works on body organs and leads to the final state of death. It works for some years on human organs even after coming round from it. It exists in the human body.This disease is not confined now in the congested town area only, but it has broken out in full swing in the rural area. We aim is to identify the factors which are the causes of the origin of dengue and its spread over society at such a large scale. It is also our aim to find the areas of society; on which consistent endeavor will help to confine in or diminish its effect in the 0- level. Information is collected on at random survey basis, especially from peoples of dengue affected area by Questionnaire Method. Intelligence is also gathered from hospital and Internet to collect data which help to indicate factors performed heavily in which situation of Society. We reached the conclusion by experiment worked in the past- information and present data.

1. Introduction

In this time, Dengue Disease is most dangerous and creating panic in the human society. The wideness of this disease not only in India, but it propagated in developing people is being affected in India whereas 2.5 billion people are involved throughout the world. As per reports of WHO, approx. 75% of people who are affected by dengue fever (DF), are belonging in the South-East region and the western pacific region. The most significant burdens of economic, social, health, and illiteracy, are the main reason for spreading out this disease. The population density of the few states like West Bengal, Uttar Pradesh, Maharashtra, Tamil Nadu of India is very high, and is also a critical reason to spread the dengue in the human community. DF is not seasonal epidemic fever, but its prevalence increases in the rainy season, because this season is advantageous for fertilization of the Aedes mosquito, which is spreading the dengue virus. This virus is related to the family of Flaviviridae, and it has four serotypes which are DENV1, DENV2, DENV3, and DENV4, are found in infected female Aedes mosquito. These species is found from 35 degrees north to 35 degrees south latitude below an altitude 1000 meters.

Four types of Dengue infection, i.e., DHFI; DHFII; DHFIII; DHFIV, are observed. Another observation is to track out the day of decrement of fever is called day-0. The day-0 date is the ticklish date of dengue affected patients wherein our research area, authors are trying to find out the symptoms after and before affecting dengue patients. After recovery of dengue, it is also observed, that the organs are damaged in the human body of the patients. All this data has accumulated in all corners of the state of west Bengal. The researchers have tested data DF, which has an 87% accuracy level.

Data Mining (DM) is a penetrating the computational trial of big data set by using an amalgamation of statistical analysis, database technology, and Machine learning with the objective of detecting the trends for getting results of research of various fields.

The authors have dedicated to applying different types of algorithms concerned with classification, clustering as well as prediction. Originators have used various kinds of methods like Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Logistics Regression (LR), Naïve Bayes (NB) and computed the result using statistical method Cohen's Kappa and ROC_AUC to maintain the level of accuracy; in this research. In this paper, the writers examined the raised frame of observation on databases anticipated to epidemic diseases like dengue.

2. Review Literature

According to previous studies, symptoms like rash, joint pains, metallic taste, headache, vomiting, in people effected by dengue can be seen within two weeks. Patients of dengue die mostly because the disease is not diagnosed quickly. Researchers have been known to use a dataset that has 58 listed ailments or diseases. This was designed following the mechanics of Bayes Server (BS), a technique of machine learning that is able to detect dengue hemorrhagic fever (DHF); by understanding the symptoms that can have any kind of relations to fever and the like. A 99.84% accurate data was found out after this result analysis.

The research work of (Fathima and Manimegalai), show how SVM (support vector machines) and data mining computational analytics can be used to track down this arboviral disease – of dengue. It could also be used to identify rules and patterns for decision-making about the work of the future, implementing the sets of data. When such a result gives the best accuracy, it becomes quite satisfactory. However, since there are more parameters, it takes a lot of time to process this. The time it needs to calculate is more. There are multiple sources from which the data is brought together. However, the dynamic process is what processes the data.

Efficient analytical methodology is extremely necessary 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 analyse 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. There is a need to recognize the redundant and inappropriate attribute, because these may act as a noise and outlier to turn to slow the processing task. These attributes may play an unfavorable effect on the perfection of the classifier, and Statistical methods will be useful to recognize these attributes. 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.

Previous studies have identified 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 researcher, 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.

Author (Hasan et al.), have described the dengue virus as 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 between 1963-1964. 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.

3. Methodology

The data mining technique is used to quiddity for necessary intelligence from clinical data to take measure evidence for the medical decisions, making symptoms for dengue patients. The data and including attributes are playing a critical role to in achieving the success in any data mining research. In this study, LR; RF; DT; SVM; NB; the classification model are used based on the confusion matrix to establish the association between real attributes and predicted class attributes. And classification model is mainly used for calculation the right and wrong classification for individual possible vale of variables. Specially, ROC_AUC, and Cohen's Kappa statistical methods are used to get the predicted result. Three performance measurements, i.e., accuracy, sensitivity, and specificity, are used in the study of research.

3.1. Support Vector Machine (SVM)

SVM is one of the classification models was invented by Vapnik et al. in 1998. To handle soft-margin SVM problems (non- linear), the researcher may try to get the maximum margin hyperplane to measure a robust separator for the puzzle classes.

Author (Chang), studied 100 samples of dengue patients, and with the support of the SVM algorithm, the ranking of the weight of dengue symptoms were found, and it helped to detect the highest weight parameters from 24 number of support vectors. Author (Gomes and Lisa), applied the SVM algorithm from gene expression data, they analyzed 12 genes of 28 dengue patients during severe viral infection. Author (I Nordin), used the Kernel function of SVM to handle instances of the relationship between the dependent, and independent variables for achieving better performance to predict dengue cases.

The equation of a linear SVM can be written as

g(x)= $\sum_{i=0}^{n} \beta i\ pi\ qit\ + \alpha 0$ ....(1)

where, qi is the instant with label pi, β is Lagrange multiplier and $\alpha 0$ is bias

The equation for Kernel SVM as

g(x)= $\sum_{i=0}^{n} \beta i\ pi\ \ k(qi.\ q).q\ + \alpha 0$ ....(2)

where, n denotes the number of support vectors and k(qi. q) is the kernel function

3.2. Decision tree (DT)

DT; model is applied in data mining to get conclusions from observation of the data set. In this model, the supervised classification tree has leaves, and branches represent class levels and conjunctions feathers. A decision and decision-making both are explicitly 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. In Classification Tree analysis, the vitiated result is the class from the data set, and in Regression analysis, vitiated result may be deliberated a real number. The most advantages of the DT are:

- The result of the observation may be bloomed graphically.

- To handle numerical and categorical transferred to 0 – 1 values.
- No need to normalize the data, only small data preparation.
- While the box model can be used in the DT, the explanation for conducting is easy and smoothly explained by Boolean logic rather than black-box model, where the result is difficult to understand for the explanation.
- Works with extensive data set.

(Rosid et al.), conducted testing on the data, using the DT; method ID3 algorithm, in terms of symptoms that affect DHE; and achieved an accuracy value above 82%.

3.3. Logistics Regression (LR)

LR; is a statistical method which has two possibilities may be true or false, i.e., 0 or 1.LR model has two categories, one is multinomial logistics has more than two outputs, and other is ordinal LR; the output depends on its input, not depend on the statistical classification.

In Regression Analysis, LR; is a continuous and categorical variable for prediction. This will be taken from the Bermouli trail that is the case of binomial on the dependent variable of a Bermoulli trail.

In the LR; method, logistic function which is cumulative distribution function of logistic distributors is used to measure the probability to maintain the relationship between categorical dependent variable or one, and more than one independent variable. In a previous study, LR; analysis was used to detect symptoms, physical signs that classified the DF; from fever related infection within first two days of affection with 74% sensitivity and 79% specificity. In their work (Chien et al.) , used of LR; to determine valid predictor variables of DF; when the probability of Type 1 error was less than 0.05.

3.4. Naive Bayes (NB)

NB; model is created based on the 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

$\mathrm{p}(\mathrm{A}_i/\mathrm{a})=\frac{p\left(\frac{a}{Ai}\right)p(Ai)}{p(a)}$ ....(3)

Where A and a are events, P denotes probability, and P(a)!=0

By definition, we may be used Chain rule in conditional probability for posterior class

$\mathrm{p}(\mathrm{a}_1,\mathrm{a}_2,\ldots.,\mathrm{a}_n/\mathrm{A}_i)=\mathrm{p}(\mathrm{a}_1/\mathrm{A}_i)\mathrm{p}(\mathrm{a}_2,\ldots..,\mathrm{a}_n/\mathrm{A}_k)$ ....(4)

By repeated:

$\mathrm{p}(\mathrm{a}_1,\mathrm{a}_2,\ldots,\mathrm{a}_n/\mathrm{A}_i)=\mathrm{p}(\mathrm{a}_1/\mathrm{A}_i)\mathrm{p}(\mathrm{a}_2/\mathrm{A}_i,\mathrm{a}_i)\ldots\mathrm{p}(\mathrm{a}_n/\mathrm{A}_i.\mathrm{a}_1,\mathrm{a}_2,\ldots$ $,\mathrm{a}_{n-1})$ ......(5)

And Conditional independence may be (assumption)

$\mathrm{p}(\mathrm{a}_1,\mathrm{a}_2,\ldots,\mathrm{a}_n/\mathrm{A}_i)=\mathrm{p}(\mathrm{a}_1/\mathrm{A}_i)\mathrm{p}(\mathrm{a}_2/\mathrm{A}_i)\ldots\mathrm{p}(\mathrm{a}_n/\mathrm{A}_i)=\prod_{i=1}^{n} p\frac{a_j}{A_i}$ ....(6)

So, posterior class probability (according to NB classifier) is

$\mathrm{p}(\mathrm{a}_1,\mathrm{a}_2,\ldots,\mathrm{a}_n\ )=\frac{1}{k}\ \mathrm{p}(\mathrm{A}_i)\prod_{i=1}^{n} p(aj/Ai)$ ....(7)

Where $\mathrm{k}=\mathrm{p}(\mathrm{a}_1,\mathrm{a}_2,\ldots.,\mathrm{a}_n)$.

Author (Caicedo-Torres, William, and Pinzón), used Gaussian Priors from NB model were implemented to find each feature mean and estimated Variance. Previous studies have used Gaussian Priors from NB model were implemented to find each feature mean and estimated Variance. Author (Arafiyah, Ria, and Hermin) , 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%.

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 classifiers 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 nonparametric, and it can handle categorical, and multimodel data which are maybe ordinal or non-ordinal. Authors have previously applied RF in combination with ANN for evaluation of dengue model performances (Silitonga and Permatasari) . They 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.

W(x)= Sign ($\sum_{i=1}^{n} \alpha i \ wi \ (x)$) ....(8)

Where, w(x) is the weight of training data, w$_i$is the weight of training data ,wi(x) refers to the output of weak classifier, I for input x and αi denotes weight operation to the classifier

α_i =0.5* log($\frac{1-E}{E}$) ....(9)

where, E denotes error rate

D$_{i+1}$(t) =$\frac{Di(t)exp(-\alpha iYt \ hi(Zt))}{pi}$ ....(10)

Where D$_i$ impliesto the weight of the previous layer. Then, the weights are normalized by dividing each of them by the sum of all weights ρi, and Yt is the y level of training point (Zt, Yt)

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= A$_0$ – A$_e$ / (1-A$_0$), where probability of agreement A$_0$ = (Number in agreement / Total)

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

A (correct) = (P + Q / P + Q + R + S) * (P + R / P + Q + R + S)

A (incorrect) = (R + S / P + Q + R + S) * (Q + S / P + Q + R + S)

Where, P The total number of raters is correct.Theraters are in agreement.

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

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

S The total number of both raters are incorrect; this is agreement.

get interpret results of rates, authors used N X N grid.

3.8. Receiver Operating Curve and Compute Area Under (ROC_AUC)

ROC; stands for Receiver Operating Curve, and AUC stands for Compute Area Under, as it compares of two operating characteristics, True Positive Rate (TPR), and False Positive Rate (FPR), that's why ROC; is also known as the relative operating characteristic curve. ROC; curve has constructed by marking the TPR; which 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 deliberated using roc_auc_score()function. Both the roc_curve and the AUC function take true outcomes, i.e. (0, 1), and enumerated the class 1.

Sensitivity or hit rate or TPR = Tp / P = Tp / (Tp + Fn) = 1 – FNR

Specificity or Selectivity or True Negative Rate (TNR) = Tn / n = Tn / (Tn + Fp) = 1 – FPR

Precision or Positive Predicted Value (PPV) = Tp / (Tp + Fp) = 1 – FDR,

Ngative Predicted Value (NPV) = Tn / (Tn + Fn),

False Negative Rate (FNR) or Miss Rate = Fn / p = Fn / (Fn + Tp) = 1 – TPR

FPR = Fp/n = Fp / (Fp + Tn) = 1 – TNR (True Negative Rate)

False Discovery Rate (FDR) = Fp / (Fp + Tp) = 1 – PPV

False Omission Rate (FOR) = Fn / (Fn + Tn) = 1 - NPV

Accuracy (ACC) = (Tp + Tn) / (p +n) = (Tp + Tn) / (Tp + Fp + Tn + Fn)

F1 Score is known as harmonic mean of precision

TABLE 1. Calculationof Cohen's Kappa

Rater-2	
Correct	Incorrect
P	Q
R	S
Correct	Incorrect
Rater-1	

and sensitivity,

So F1 = 2(PPV.TPR) / (PPV + TPR) = 2Tp / (2Tp + Fp + Fn)

Where Tp = True Positives, Tn = True Negatives, Fp = False Positives, Fn = False negatives.

4. Data sets

The data in samples of dengue patients of all over the West Bengal, were collected not only from several Hospitals but also interacted with the people individually were suffered from DF by using questionnaires method. The data containing the patient's information was diagnosed by the researcher from the year 2016 to 2019.

4.1. Questionnaires for Survey

- Patient details:

1. Name of the Patient:
2. Name of Village / Town / District:
3. Mention the residence is under Panchayat Area or Municipality area:
4. Age:
5. Gender: M/F
6. No of educated people in family:
7. Occupation:
8. What precautions you have taken against mosquitoes bite?

- What are the symptoms before detecting Dengue?
- How much temperature level is increased?
- How many days stayed Fever?
- What are the Symptoms after detecting Dengue?
- Observation of Plate late counting
- Mention Blood pressure is fluctuating or not (mention BP):
- Whether the Bleeding happened?
- What are the symptoms getting after recovery from Dengue?
- Have you been admitted in the Hospital / Nursing home?
- What are the medical treatments, tests, etc. have taken in the hospital?
- Do you have any effect on any other organ after Dengue?
- What are the key reasons for effecting Dengue?

During the study, the clinical statement was recorded from the patient at various stages, and the data consists of 91 patients and 13 fields. All this collected raw data does not consider in the experiments. These fields have carried a significant role in the study. Three area – Symptoms after and before the detection of DF and third is, what body organ/organs are affected in the human body of the patient, after recovery from DF.

Authors have taken 13 fields name are: age, symptoms before detecting dengue, test, temperature level, fever stayed in days, symptoms after detecting dengue, plate late counting, average bp, bleeding happened or not, symptoms after recovery of dengue, hospitalized or not, affect other organs after recovery.

5. Results

Number of True cases: 41 (45.05%)

Number of False cases: 50 (54.95%)

12% in training set

88% in test set

Original True : 41 (0.00%)

Original False : 50 (0.00%)
Training True : 30 (0.00%)
Training False : 42 (0.00%)
Test True : 11 (0.00%)
Test False : 8 (0.00%)

5.1. Creating new classifier with ensemble

After testing with ensemble of LR, RF, NB, DT individually using AdaBoost model we get the given below result.

From the report, it is observed, that highest accuracy level is reached for using RF with AdaBoost machine learning method. We proposed to call this new combination Ensemble Random Forest (ERF) Model. This is our newly designed classifier which appears to be most effective in this particular application. The summarized report using ERF is given below**.**

5.2. Summarized report using ERF Model

Accuracy=81.556
Standard deviation=0.12374805148734092
Sensitivity=82.889
Precision=84.606
F Score=83.739
ROC_AUC:87.15
Kappa Score:68.908

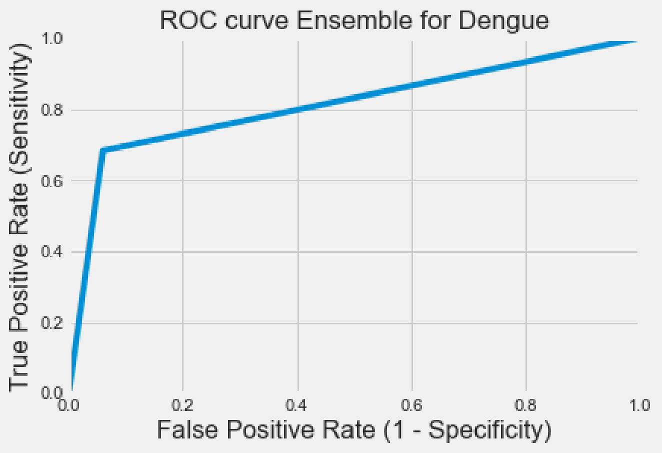

FIGURE 1. ROC curve ensemble for Dengue (Output 1)

5.3. Ensemble New Model

Accuracy=87.0
Standard deviation=0.1342606641050383
Sensitivity=78.111
Precision=83.741
F Score=80. 828
ROC_AUC:89.25
Kappa Score:66.986

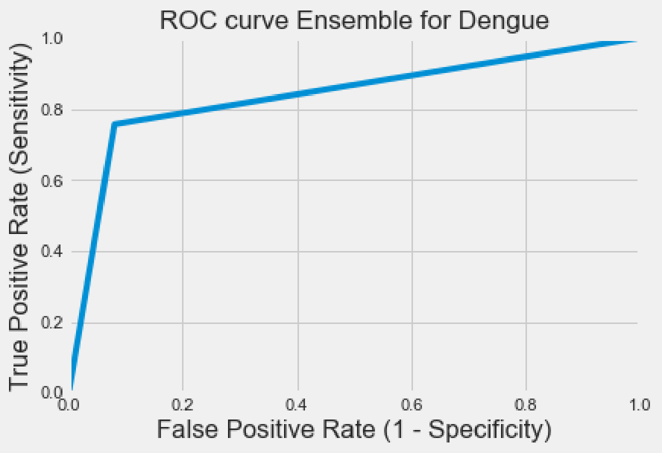

FIGURE 2. ROC curve ensemble for Dengue (Output 2)

The Author (Rohan and Islam) used the same ensemble of RF with AdaBoost was used for detecting of Brest Cancer. They analyzed on 699 instances, where 458 of benign data, 241 of malignant data, 11 features, and 10 attributes. In the testing phase; structured provided 98.5714% of accuracy, Sensitivity 100%, and specificity 96.296%.

The introduced model performs better than conventional RF classifier.

6. Discussions of Results:

We have analyzed 91 numbers of patient's information where 79.12% training set and 20.88% test set, and at last, got the results from a different aspect of using several models are summarized.

From the data set classified by LR; and Confusion matrix, accuracy is 0.4737. Using Cohen's Kappa Score is 0.020619 and ROC_AUC: 0.511364

Using RF; from Confusion matrix, the Training accuracy is 0.9722, and the testing accuracy is 0.7368, Cohen's Kappa Score is 0.486486 and ROC_AUC: 0.755682

Training accuracy, and the Testing accuracy for the NB; model are 0.4722 and 0.6316, respectively. The Score of Cohen's Kappa is 0.14195 and ROC_AUC is 0.562500

For DT; Training accuracy is 1.0, Testing accuracy is 0.8421 and Cohen's Kappa: 0.681564 and ROC_AUC: 0.84

For SVM from confusion matrix, Training accuracy is 0.7083 and Testing accuracy is 0.3684, Cohen's Kappa = 0.40 and ROC_AUC = 0.420435

TABLE 2. Classification report on the basis of precision

	LR	RF	NB	DT	SVM
Micro Average	0.47	0.74	0.63	0.84	0.37
Macro Average	0.51	0.76	0.81	0.84	0.35
Weighted Average	0.53	0.77	0.77	0.85	0.35

TABLE 3. Classification report on the basis of recall

	LR	RF	NB	DT	SVM
Micro Average	0.47	0.74	0.63	0.84	0.37
Macro Average	0.51	0.76	0.56	0.85	0.42
Weighted Average	0.47	0.74	0.63	0.84	0.37

TABLE 4. Classification report on the basis of f1 score

	LR	RF	NB	DT	SVM
Micro Average	0.47	0.74	0.63	0.84	0.37
Macro Average	0.46	0.74	0.49	0.84	0.32
Weighted Average	0.45	0.74	0.53	0.84	0.29

TABLE 5. Classification report on the basis of support

	LR	RF	NB	DT	SVM
Micro Average	19	19	19	19	19
Macro Average	19	19	19	19	19
Weighted Average	19	19	19	19	19

TABLE 6. Statistical report with ensemble

	LR	RF	NB	DT	SVM
Cohen's Kappa	0.020619	0.486486	0.141935	0.681564	0.140006
ROC_AUC	0.511364	0.755682	0.562500	0.846591	0.420455

TABLE 7. Classification report on the basis with ensemble of precision

	LR	RF	NB	DT
Micro Average	0.42	0.74	0.47	0.84
Macro Average	0.45	0.73	0.55	0.84
Weighted Average	0.46	0.74	0.57	0.85

TABLE 8. Classification report on the basis with ensemble of recall

	LR	RF	NB	DT
Micro Average	0.42	0.74	0.47	0.84
Macro Average	0.47	0.74	0.53	0.85
Weighted Average	0.42	0.74	0.47	0.84

TABLE 9. Classification Report on the basis with Ensemble of f1 Score

	LR	RF	NB	DT
Micro Average	0.42	0.74	0.47	0.84
Macro Average	0.59	0.73	0.43	0.84
Weighted Average	0.37	0.74	0.41	0.84

TABLE 10. Classification report on the basis with ensemble of support

	LR	RF	NB	DT
Micro Average	19	19	19	19
Macro Average	19	19	19	19
Weighted Average	19	19	19	19

TABLE 11. Statistical report with ensemble

	LR	RF	NB	DT	SVM
Cohen's Kappa	0.060914	0.469274	0.500000	0.681564	0.060914
ROC_AUC	0.4659090	0.738636	0.528409	0.846591	0.4659090

After ensemble, the RF; the result of the prediction accuracy has increased up to 0.7368

Finally, experimented on Ensembled new model, resulting accuracy level is reached up to 87.0, Standard Deviation = 0.134, Precession = 83.74, Sensitivity =78.11, F Score: 80.828, ROC_AUC = 89.25 and Kappa Score: 66.986

RF; a single classifier used on data to measure the performance shown in table-1. They did not investigate the plurality of the number of classifiers such as LR, SVM, DT, NB, whereas, we used all stated model for proper investigation in this paper and all respect of the above mentioned models, minimizing FPR; and FNR; and at the end, got the predicted result. A. Osarumwense, B. Eromosele, 2020, implemented, the popular machine-learning Bayesian Belief network model was designed on the Bayes Server platform to predict Hemorrhagic Fever and its symptoms. In comparison to others (Balasaravanan and Prakash), our prediction accuracy much better is shown in the table-1. Researchers, Dasgupta, Sharma, Sinha, Raghavendra, 2019, used three machine learning algorithms RF; DT, and SVM, on their survey data, and they found accuracy level much better shown in the below table-1. (Mello-Román et al.), worked on 90% training and 10% testing data of data set of early detection, and diagnosis of DF. They used Artificial Neural Network (ANN), where multilayer perception (MLP), and Radial basis function (RBF) were applied, and SVM classifier, where three kernel function have been evaluated with the support of IBM; SPSS; software. Using of the ANN-MLP classifier, they gained 96% accuracy, 96% sensitivity and 97% specificity with low validation, and using of SVM – Polynomial, got result of 90% for accuracy level. In the research of (Salami), for predetection of DF; four machine learning models (pls, glmset, RF; xgboost) were evaluated with testing data set ROC_AUC as the quantitative measure for performance measure shown in the given table-1.

In the paper of (Kapoor, Kadyan, and Ahuja), the authors have conducted an analytical study conducted and used standard parameters, and prepared a dataset for a machine learning predictive model to detect DF, for early detection. The main target of that paper was four main factors which are fever, skin disease, headache, abdominal pain for early detection of dengue.

7. Conclusion

From the table-1, it has observed that, accuracy label result is more affective, i.e., 87.0% using several machine learning models as stated above, and two most effective statistical methods Cohen's Kappa, and ROC_AUC. From the analysis of data, symptoms of the dengue patients before, and after detection of DF are specifically marked. And very important conclusion in the paper, it has observed that, body parts have been damaged in the patient's body after recovery from DF, This damage may be minor and major significant effect in the human body. Damage organ may be Liver, Prostate, Spleen, and different Spot shown on the human body, Enzyme system failure, the Nerve system failure in the brain and also from the study of data, the dengue patients suffers from blood sugar, weight loss, appetite and weakness after recovering from DF. But in our study, it reveals that, four factors are not sufficient to detect dengue; these, four factors are now the similar symptoms to other diseases like COVID-19. The author has looked that, more symptoms have added from their battue. Eventually, it has concluded by the author that, except the above four factors, eye pain or red eye or both, hiccups, and loose stool are

TABLE 12. Quantitative measure for performance measure of four machine learning models (pls, glmset, RF; xgboost)

In Paper	Accuracy	Standard Deviation	Sensitivity	Specificity	Precision	F_Score	ROC_AUC	Kappa Score
1	-	-	88.1%	94.9%	-	-	-	-
2	92.34%	-	94.04%	92.19%	-	-	-	-
3	-	-	80%	65%			0.75	
4	90%	-	96%	97%	-	-	-	-
5	95.00%	-	-	-	-	-	-	-
6	88%	-	-	-	-	-	0.94	-
Our Study	**87.0%**	**0.13426066 41050383**	**78.111**	**-**	**83.741**	**80.828**	**89.25**	**66.986**

also added as major common factor for early detection of DF.

8. Future Scope

The proposed technique has been tested only on dengue classification, and it should further have evaluated clinical datasets. The proposed methodology can be tested on other applications where nature of data is different.

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 (2018). 10.1088/1742-6596/948/1/012077.

Balasaravanan, K. and M. Prakash. “Detection of dengue disease using artificial neural network based classification technique”. *International Journal of Engineering &amp; Technology* 7.1.3 (2017): 13–13. 10.14419/ijet.v7i1.3.8978.

Caicedo-Torres, Ángelpaternina William, 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.

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.

Chien, et al. “An app detecting dengue fever in children: using sequencing symptom patterns for a web-based assessment”. *JMIR mHealth and uHealth 7.5* (2019). 10.2196/11461.

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.

Gomes, Ana and V Lisa. “Classification of dengue fever patients based on gene expression data using support vector machines”. *PloS one* 5 (2010). 10.1371/journal.pone.0011267.

Hasan, Shamimul, et al. “Dengue virus: A global human threat: Review of literature”. *Journal of International Society of Preventive and Community Dentistry* 6.1 (2016): 1–1. 10.4103/2231-0762.175416.

I Nordin, N. “The Classification Performance using Support Vector Machine for Endemic Dengue Cases”. *Journal of Physics: Conference Series* 1496.1 (2020). 10.1088/1742-6596/1496%20/1/012006.

Kapoor, Rajeev, Virender Kadyan, and Sachin Ahuja. “Identification of Influential Parameter for Early Detection of Dengue Using Machine Learning Approach”. *Proceedings of the 5th International Conference on Cyber Security & Privacy in Communication Networks (ICCS). 2019* (). 10.2139/ssrn.3511419.

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.

Rohan, Tanbin and Islam. “A precise breast cancer detection approach using ensemble of random forest with AdaBoost”. *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)* (2019). 10 . 1109 / IC4ME % 20247184 . 2019 . 9036697.

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.

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

Silitonga and Permatasari. “Evaluation of Dengue Model Performances Developed Using Artificial Neural Network and Random Forest Classifiers”. *Procedia Computer Science 179* (2021). 10.1016/j.procs.2020.12.%20018.

Embargo period: The article has no embargo period.

To cite this Article: Bhattacharya, Kousik, Dr. Avijit Kumar Chaudhuri, Anirban Das, and Dilip K Banerjee. “**Comparison of Recent Data Mining Algorithms to Identify of the factors and effects of Dengue Fever and Ensemble Random Forest, A new Algorithm**.” International Research Journal on Advanced Science Hub 04.05 May (2022): 143–153. http://dx.doi.org/10.47392/irjash.2022.034