CrossRef DOI of original article:

Clinical Dengue Data Analysis and Prediction using Multiple Classifiers: An Ensemble Techniques

Veena Kumari H M[1], Dr. uresh D S[2] and Dr. Dananjaya P E[3]

[1] Channabasaveshwara Institute of Technology, Gubbi,Tumkur

Received: 1 January 1970 Accepted: 1 January 1970 Published: 1 January 1970

Abstract

The dengue infection is caused by the mosquito Aedes aegypti. According to WHO, 50 to 100 million dengue infections will occur every year. Data-miming techniques will extract information from the raw data. Dengue symptoms are fever, severe headache, body pain, vomiting, diarrhea, cough, pain in abdomen etc. The research work is carried out on real data and the patient data is collected from the Department of General Medicine, PESIMSR, Kuppam, Andrapradesh. Dataset consists of 18 attributes and one target value. Research work has been done on binary classification to classify dengue positive (DF) and dengue negative (NDF) cases using different ML techniques. The proposed work demonstrates that ensemble techniques bagging, boosting and stacking gives better results than other models. The Extreme Gradient Boost (XGB), Random Forest by majority voting and stacking with different meta classifiers are the ensemble techniques used for the binary classification. The dataset is divided into 80

Index terms— dengue fever, aedes aegypti, XGB, stacking, ROC, AUC.

1 I. Introduction

engue fever (DF) is an arthropod-borne viral disease common past three decades. According to WHO, 51-101 million new infections with dengue occur every year in more than a hundred endemic countries [1]. Dengue fever is a severe viral infection with potentially fatal consequences. Dengue fever was originally known as "water poison." The dengue caused by the female Aedes aegypti mosquito is shown in Fig. **??** A Female Aedes Aegypti Mosquito

In the 1780s, the first clinically recognized epidemics of dengue occurred at the same time in Africa, Asia, and North America. Benjamin Rush was named "break-bone fever" based on the features of arthralgia and myalgia. The dengue epidemic was first reported in Chennai in 1780, the first virologically proven outbreak of dengue fever in India appeared at Calcutta and the East Coast of India in 1963-64. In the 1970s and 1980s, epidemic activity accelerated dramatically, resulting in the widespread of viruses and mosquito vectors and the consequent DENV transmission across the world [2]. The first major DHF epidemic occurred in the Philippines during 1953-1954, continued by a rapid global spread of DF/DHF epidemics. The first major DHF/DSS epidemics in India occurred in 1996, at Delhi and Lucknow, and later extended throughout the country. In India outbreaks of dengue have become more common in many parts. Between 2010 to 2014 incidence of reported cases of dengue was 34.81 per million population. Dengue fever became endemic in Orissa, Uttarakhand, Bihar, Assam, and Jharkhand, in 2010 [3].

2 II. Background Study

Kassaye Yitbarek Yigzaw et al [2] presented a benchmarking platform for the prediction of communicable diseases. Rathi et al [4] studied dengue infection in Rajasthan. The study was based on 100 admitted children and he classified the patients based on their symptoms. Kalayanarooj S [3] demonstrates the clinical appearances of dengue and DHF. Aldallal, A.S [5] explained that data mining techniques are used for the prediction of non-communicable diseases like heart and diabetes. Agrawal et al [7] demonstrated the ensemble approach by using multiple classifiers Ada boost, and a decision tree for the prediction of diabetes. Ghosh et al [10] used multiple classifiers for the sentiment analysis performance assessment. Gupta et al [12] compared different ML approaches for heart disease prediction. Mesafint et al [14] explained ML algorithms for the prediction of HIV/AIDS tests.

3 III. Proposed Methodology

The ensemble models are Extreme Gradient Boost (XGB), Random Forest (RF) by majority voting, and Stacking, which is based on a combination of heterogeneous classifiers like NB, KNN, and SVM. It is very helpful to consider ensemble techniques [6], for dengue fever diagnosis and prediction. The proposed framework is shown in Fig 3. The main aim of data acquisition and the data pre-processing module is to get the Dengue fever dataset and process them into a suitable form for further analysis. Datasets have features/attributes which will finally distinguish the data into patient sick and healthy. The dataset has thirty-eight features and different data types. The dataset is spitted into an 80% training set and a 20% testing dataset. The pre-processing includes feature selection and missing value imputation [8]. The proposed model combines different classifiers such as Naïve Bayes, K -Nearest Neighbor, and Support vector machine. For each classifier, the output is predicted.

Each base classifier is used in the ensemble framework by training data to make it useful for the prediction of dengue. Dataset features and target values are known to each classifier, which in turn can predict whether the disease is present or not.

4 i. Description of the Dengue Dataset

The patient data is collected from the Department of General Medicine, PESIMSR, Kuppam, Andrapradesh. The patient is diagnosed in the laboratory using the dengue duo card test shown in fig 4. Dataset consists of 18 attributes and one target value. The number of patients having each symptom is listed in Table I and corresponding bar charts explain the importance of each feature [9] are shown in fig. **??**. Among 140 dengue-infected cases all the patients are suffering from fever,106 headache, 97 and 94 myalgia and arthralgia and 83 low back pain and others.

5 ii. XGBoost

Boosting is a broadly used and highly effective machine learning algorithm. An end-to-end tree boosting system called XGBoost is widely used by data experts. The important factor is its scalability for better accuracy. The system is ten times faster than existing conventional methods. The scalability of XGBoost is due to several algorithm optimizations. Parallel and distributed computing will make learning faster [15]. In the stacking algorithm, the base (first-level) classifiers are trained by the same set of the training sample, which is used to prepare the inputs for the meta (second-level) classifier, which may cause overfitting. The stackingCVclassifier uses the cross-validation method. The dataset is split into k folds, and k-1 folds are used to fit the level-1 classifier in k successive rounds. In every iteration, the level-1 classifiers are then applied to the remaining subset. The predictions of the base classifiers are then stacked and which is an input to the level-2 classifier.

6 NO. OF PATIENTS

7 IV. Performance Evaluation

The clinical dengue fever data set was used to analyse the performance of the ensemble model and to compare it with the other models. The class labels dengue infected (DF) with the dengue not infected (NDF) is replaced with class 1 and class 0 to maintain uniformity [16]. Each dataset is split into training and testing sets. Cross validations of 10-fold are applied. performance measure of each base classifier, as well as the ensemble model, is calculated using a confusion matrix. The base classifiers NB, SVM & KNN are trained first and then they are tested. The proposed research work analysed the performance of the ensemble methods XGB, RF, and Stacking. The metrics are accuracy, recall, precision, and f1-score. The confusion matrix illustrates the actual and predicted classification [15,17]. The equations (**??**), (**??**), (3), and (**??**), are used to calculate the metrics [17]. III and Fig. **??**1. The ensemble methods XGB, RF, and Stacking give 98.57%, 99.12%, and 99.56% for the training dataset, whereas 97.80%, 94.82% and 98.27% for the testing dataset. We observed better accuracy for ensemble methods. IV. The AUC for the proposed ensemble XGB is 97.14% and 97.81% for random forest 98.14% and 99.14%, for stacking 98.14% and 98.68% for testing and Training datasets respectively. As shown in Table III, the AUC values for the datasets lie between 0.97 to 0.99, indicating that the positive class values are correctly distinguished from the negative class values.

8 Table II: Confusion Matrix

9 Actual

10 Table V: Auc Comparision

11 V. Conclusion

The main objective of this research work is to the prediction of dengue fever using ensemble techniques. We used bagging, boosting, and stacking methods for prediction and the end results are compared with the NB, KNN, and SVM models. The experimental results prove that Ensemble techniques are the best models for the prediction of dengue fever. The techniques were analysed using performance metrics. The accuracy for the extended boost, random forest with majority voting, and stacking using metaclassifiers gives better accuracy for both the training and testing datasets compared to other models. The extended analysis was done by using the roc curve and precision-recall curve, which explains the performance of the models. The Area under the curve lies between 0.97 to 0.99. The ensemble models are the better models for the prediction of dengue-infected patients.

2

Figure 1: Fig. 2 :

3

Figure 2: Fig. 3 :

4

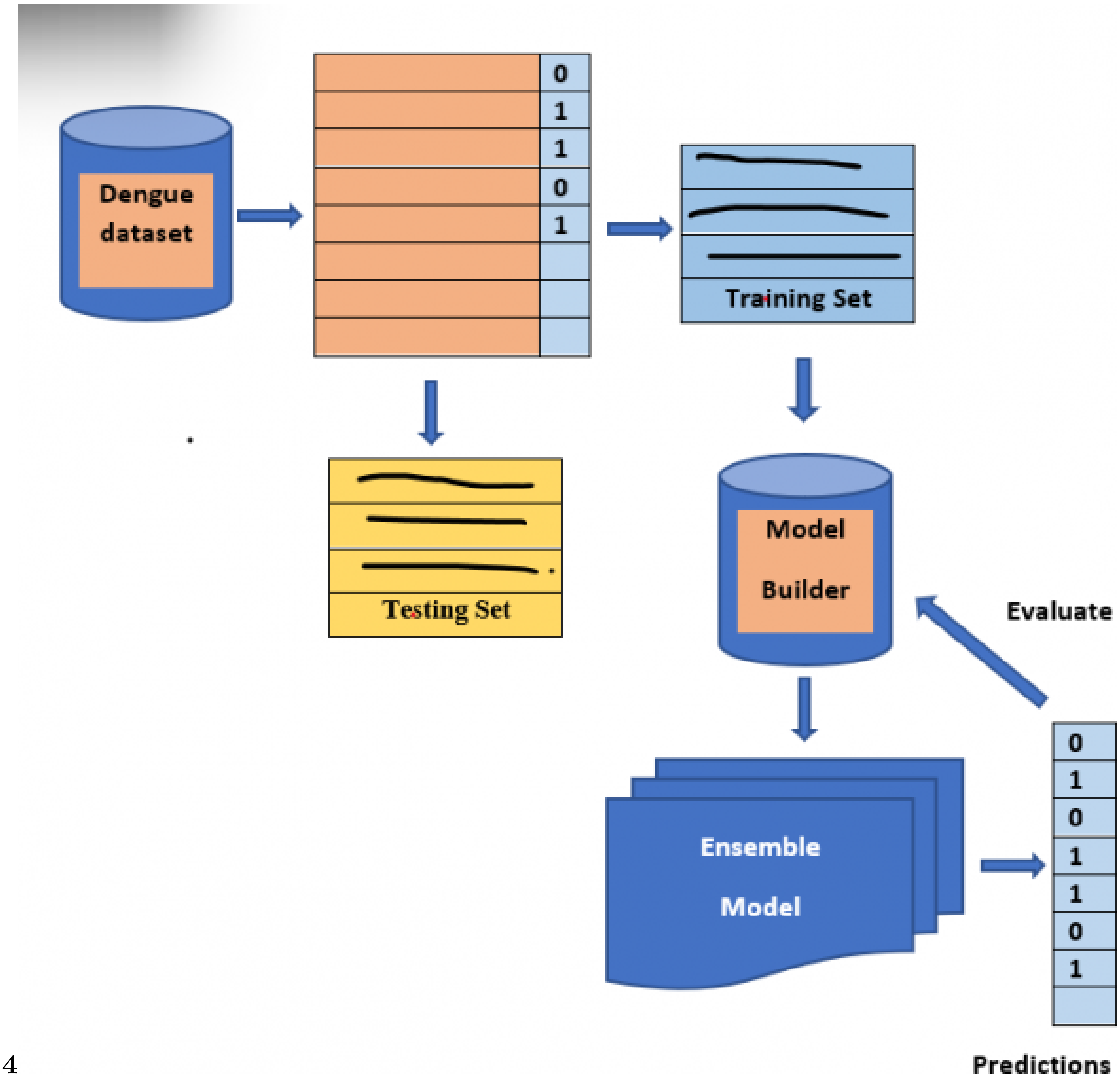
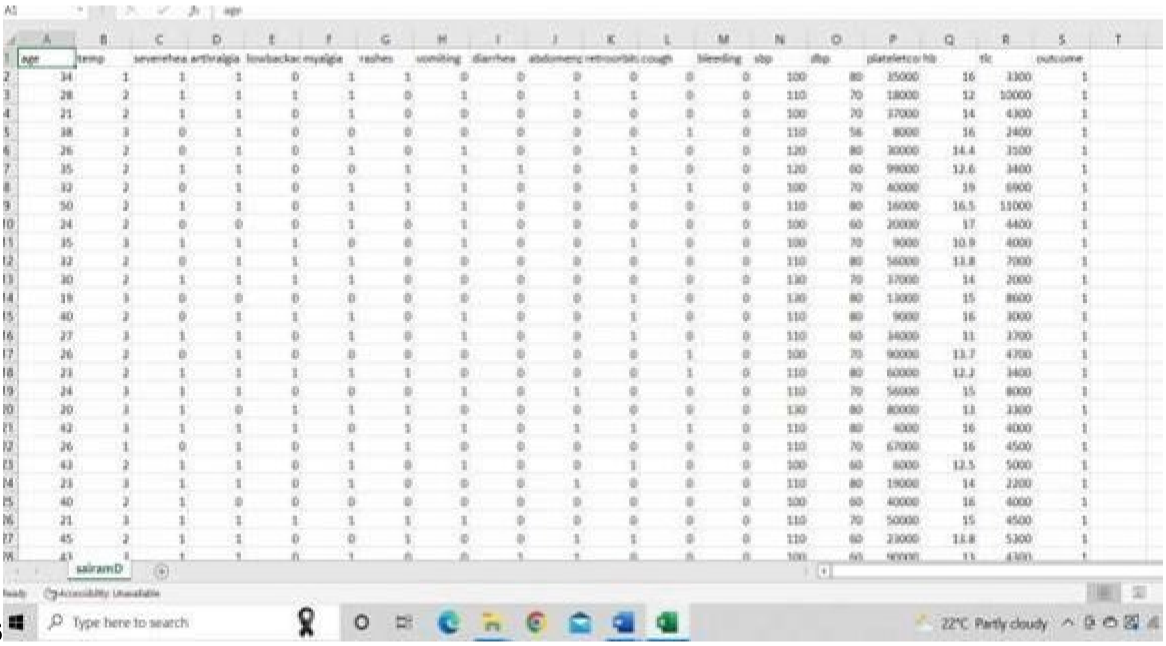


Figure 3: Fig. 4 :

5

Figure 4: Fig. 5 :



	age	temp	severehea	arthralgia	lowbackac	myalgia	rashes	vomiting	diarrhea	abdomen	retroorbit	cough	bleeding	sbp	dbp	plateletco	hb	tlc	outcome
2	34	1	1	1	0	1	1	0	0	0	0	0	0	100	80	35000	16	3300	1
3	28	2	1	1	1	1	0	1	0	1	1	0	0	110	70	18000	12	10000	1
4	21	2	1	1	0	1	0	0	0	0	0	0	0	100	70	37000	14	4300	1
5	38	3	0	1	0	0	0	0	0	0	0	1	0	110	56	8000	16	2400	1
6	26	2	0	1	0	1	0	1	0	0	1	0	0	120	80	30000	14.4	3100	1
7	35	2	1	1	0	0	1	1	1	0	0	0	0	120	60	99000	12.6	3400	1
8	32	2	0	1	0	1	1	1	0	0	1	1	0	100	70	40000	19	6900	1
9	50	2	1	1	0	1	1	1	0	0	0	0	0	110	80	16000	16.5	11000	1
10	24	2	0	0	0	1	0	1	0	0	0	0	0	100	60	20000	17	4400	1
11	35	3	1	1	1	0	0	1	0	0	1	0	0	100	70	9000	10.9	4000	1
12	32	2	0	1	1	1	0	0	0	0	0	0	0	110	80	56000	13.8	7000	1
13	30	2	1	1	1	1	0	0	0	0	0	0	0	130	70	37000	14	2000	1
14	19	3	0	0	0	0	0	0	0	0	1	0	0	130	80	13000	15	8600	1
15	40	2	0	1	1	1	0	1	0	0	1	0	0	110	80	9000	16	3000	1
16	27	3	1	1	0	1	0	1	0	0	1	0	0	110	60	34000	11	3700	1
17	26	2	0	1	0	0	0	0	0	0	0	1	0	100	70	90000	13.7	4700	1
18	23	2	1	1	1	1	1	0	0	0	0	1	0	110	80	60000	12.2	3400	1
19	24	3	1	1	0	0	0	1	0	1	0	0	0	110	70	56000	15	8000	1
20	20	3	1	0	1	1	1	0	0	0	0	0	0	130	80	80000	13	3300	1
21	42	3	1	1	1	0	1	1	0	1	1	1	0	110	80	4000	16	4000	1
22	26	1	0	1	0	1	1	0	0	0	0	0	0	110	70	67000	16	4500	1
23	43	2	1	1	0	1	0	1	0	0	1	0	0	100	60	8000	12.5	5000	1
24	23	3	1	1	0	1	0	0	0	1	0	0	0	110	80	19000	14	2200	1
25	40	2	1	0	0	0	0	0	0	0	0	0	0	100	60	40000	16	4000	1
26	21	3	1	1	1	1	1	1	0	0	0	0	0	110	70	50000	15	4500	1
27	45	2	1	1	0	0	1	0	0	1	1	0	0	110	60	23000	13.8	5300	1
28	43	3	1	1	0	1	0	0	1	1	0	0	0	100	60	90000	13	4300	1

6

Figure 5: Fig. 6 :

89

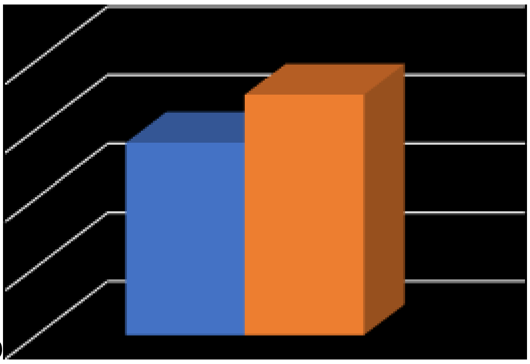


Figure 6: Fig. 8 :Fig. 9 :

Figure 7:

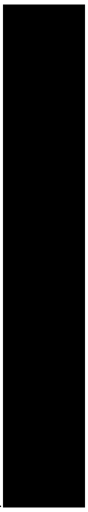
1011 

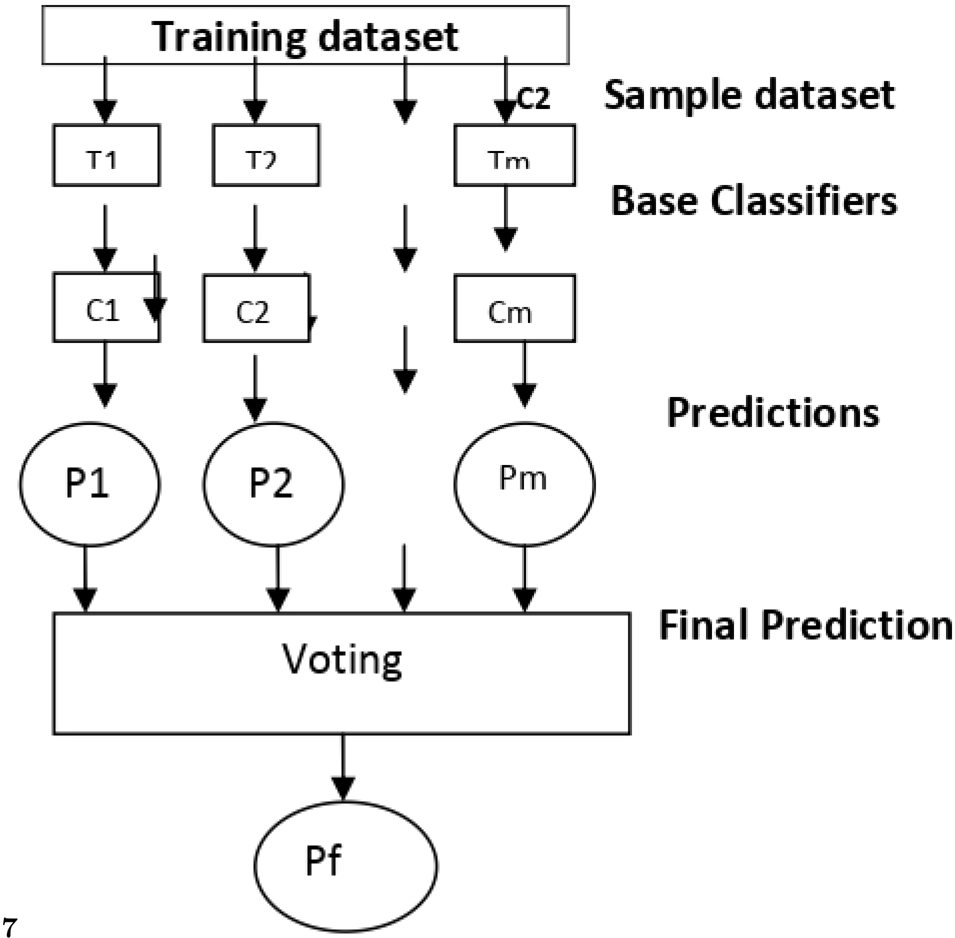
Figure 8: Fig. 10 :Fig. 11 :

13 

Figure 9: Fig. 13 :

14 

Figure 10: Fig. 14 :



1617

Figure 11: Fig. 16 :Fig. 17 :

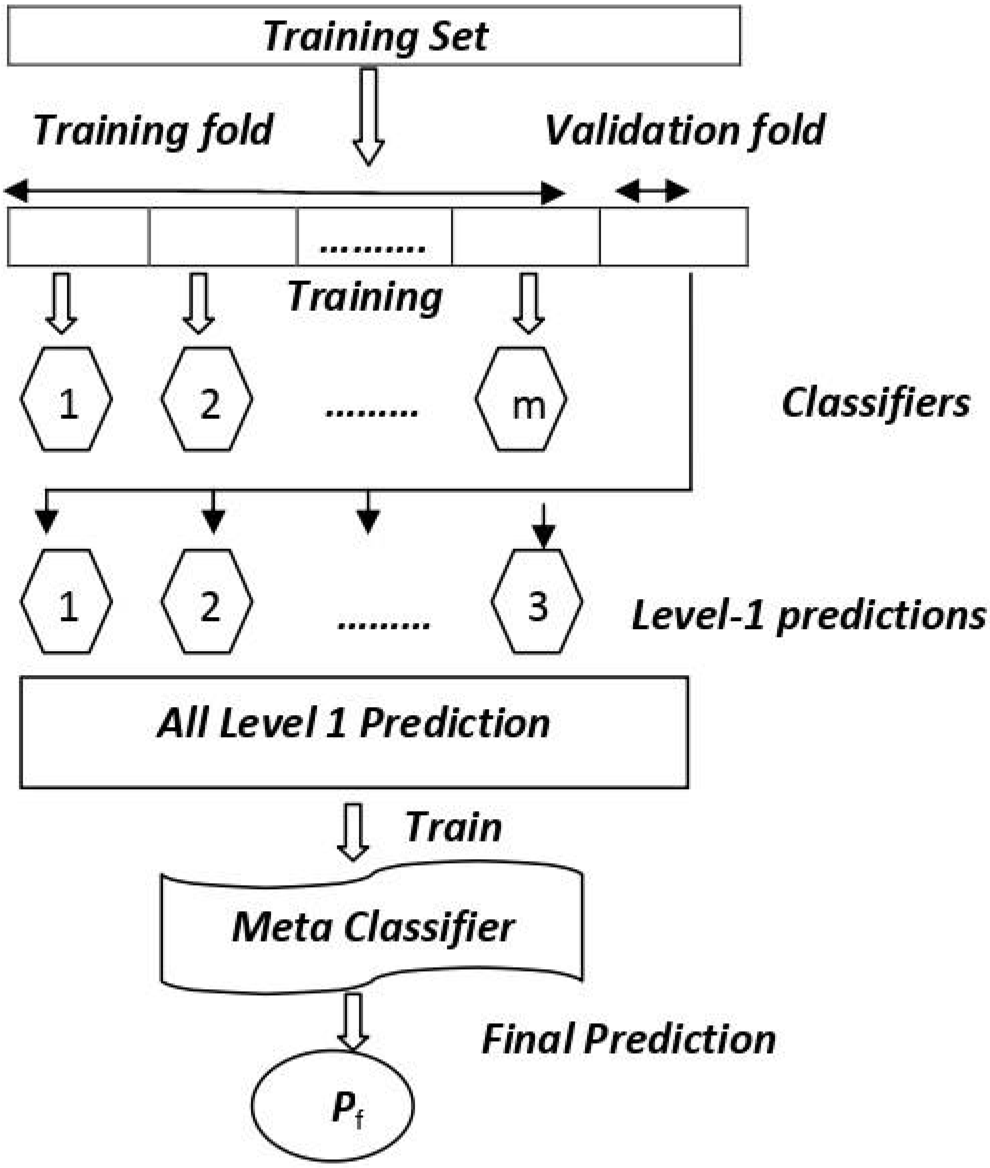


Figure 12:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.$$

Figure 13:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2)$$

Figure 14:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Figure 15:

I

Target

200
150
100 50

Fig. 7: Bar Chart Representation

b) Ensemble Methods

Clinical Feature Ensemble means combining multiple models. This approach gives better performance comp

Fever single model. Thus, a set of models is used for Headache predictions than a single model [7]. The mai

Myalgia errors. If the ensemble technique of bagging, boosting,
Arthralgia and stacking are used for classification, high accuracies
Low Backache can be obtained. Bagging creates a different subset of
Retro Orb Pain training data from the sample training dataset & the final Rashes output depends on major

XGBOOST and ADA BOOST
i.

Bleeding
Cough
Diarrhea
Sore
Throat
Breathlessnes
Seizures

Figure 16: Table I :

SVM RF Training Dataset

NB

Accuracy of Random Forest: 99.12

Accuracy of Naive Bayes model: 95.40 precision recall f1-score 0 0.93 0.98 0.9 1 0.98 0.92 0.95 Accuracy of S

XGB

KNN

Accuracy of Extreme Gradient Boost

:98.57 precision	recall	f1-score
0	0.99 0.97	0.98
1	0.97 0.99	0.98

Figure 17: Accuracy of K-Neighbors Classifier :96.49 precision recall f1-score 0 0.93 1.00 0.97 1 1.00 0.93 0.96 Accuracy of K-Nearest Neighbour : 85.66 precision recall f1-score 0 0.86 0.97 0.91 1 0.96 0.81 0.88

[Note: Matrix and Experimental Results of Training and Testing Dataset of the Ensemble and Other M Models]

Figure 18: 97 Stacking Accuracy of Stacking CV Classifier :99.56 precision recall f1-score 0 1.00 0.99 1.00 1 0.99 1.00 1.00 Stacking Accuracy of Stacking CV Classifier: 98.27 precision recall f1-score 0 0.97 0.99 0.98 1 0.99 0.96 0.98 Confusion

III

Year 2022

47

( ) D

Classifiers	Training Dataset	Testing Dataset
NB	95.40	93.17
KNN	96.49	85.66
SVM	97.51	89.65
XGB	98.57	97.80
RF	99.12	94.82
Stacking	99.56	98.27

Figure 19: Table III :

IV

110
90 97.5
80
NB KNN SVM XGB RF

Year 2022
48
Volume XXII Issue II Version I
( ) D Global Journal of Computer Science and Technology

Classifiers NB KNN SVM RF XGB Ensemble Stacking

Training dataset Precision Recall (%) (%) NDF 93 98 DF 98 92 NDF 93 100 DF 100 93 NDF 96 98 DF 99 100 NDF 98 100 DF 100 98 NDF 99 97 DF 97 99 NDF 100 99 DF 99 100

f1-score (%) 95 95 97 96 97 100 99 99 98 98 100 100

Classifiers NB KNN SVM RF XGB Ensemble

Figure 20: Table IV :

Year 2022
50
( ) D

Classifier	Testing Dataset	Training Dataset
Auc_Nb	0.9629	0.9514
Auc_Knn	0.8333	0.9342
Auc_Svc	0.9444	0.9956
Auc_Xgb		0.9781
Auc_Rf	0.9814	0.9914
Auc_Scv	0.9814	0.9868

Figure 21:

.1 Acknowledgment

Our sincere thanks to Dr. Veerapuram Manoj Reddy, Department of General Medicine, PES Medical sciences and Research, Kuppam, Andrapradesh for his support for the collection of dengue data.

[Miao and Miao ()] , Julia & Miao , Kathleen Miao . 2018.

[Yigzaw and Bellika ()] 'A communicable disease prediction benchmarking platform'. Kassaye Yigzaw , Johan Bellika . BHI2014.564-568.10.1109/BHI.2014.6864427. *IEEE-EMBS International Conference on Biomedical and Health Informatics*, 2014. 2014.

[Nynalasetti et al. ()] 'A hybrid Algorithm for Dengue Disease Prediction with Multi Dimensional Data'. Konadala Kameswara Rao & Nynalasetti , Dr G P Varma , Saradhi . *International Journal of Advanced Research in Computer Science and Software Engineering* 2014. 14 p. .

[Cardiotocographic Diagnosis of Fetal Health based on Multiclass Morphologic Pattern Predictions using Deep Learning Classifica 'Cardiotocographic Diagnosis of Fetal Health based on Multiclass Morphologic Pattern Predictions using Deep Learning Classification'. 11.10.14569/IJACSA.2018.090501. *International Journal of Advanced Computer Science and Applications* 9.

[Reddy et al. ()] 'Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction'. Chandra Reddy , N S Shue Nee , S Zhi Min , L Ying , C . 10.11113/ijic.v9n1.210. `https://doi.org/10.11113/ijic.v9n1.210` *International Journal of Innovative Computing* 2019. 9 (1) .

[Kalayanarooj (2011)] *Clinical Manifestations and Management of Dengue/DHF/DSS*, S Kalayanarooj . 10.2149/tmh.2011-S10.Epub. 22500140. PMC3317599. 2011 Dec. 2011 Dec 22. 39 p. . (Trop Med Health. Suppl)

[Veena and Suresh ()] 'Comparative Study of Classification algorithms used for the Prediction of Non-communicable diseases'. H M Veena , D S Suresh . 10.30534/ijeter/2021/14972021. *Int. J. Emerg. Trends Eng. Res* 2021. 9 (7) p. .

[Gupta et al. ()] 'Comparison of various machine learning approaches uses in heart ailments prediction'. Gunjan & Gupta , U & Adarsh , N Reddy , B Subba & Rao , Ashwath . 2161.012010.10.1088/1742-6596/2161/1/012010. *Journal of Physics: Conference Series*, 2022.

[Shaukat et al. ()] 'Dengue Fever Prediction: A Data Mining Problem'. K Shaukat , N Masood , S Mehreen , U Azmeen . 10.4172/2153-0602.1000181. *J Data Mining Genomics Proteomics* 2015. 6 p. 181.

[Agrawal et al. ()] *Diabetes Diagnosis Prediction Using Ensemble Approach*, Agrawal , G Bhargav , E Spandana . 10.1007/978-981-15-5546-6_66. 2021.

[Heidari and Hanif Hellstern ()] *Early heart disease prediction using hybrid quantum classification*, Heidari , Gerhard Hanif & Hellstern . 10.48550/arXiv.2208.08882. 2022.

[Mesafint ()] 'Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results'. Daniel & D H Mesafint , Manjaiah . 12.10.1080/1206212X.2021. *International Journal of Computers and Applications. 1* 2021. 19746. p. 63.

[Bashir et al. ()] 'IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework'. Saba Bashir , Usman Qamar , Farhan Hassan Khan . 10.1016/j.jbi.2015.12.001. `https://doi.org/10.1016/j.jbi.2015.12.001` *Journal of Biomedical Informatics* 1532-0464. 2016. Pages 185-200. 59.

[10. Ghosh, Monalisa Sanyal, Prof(Dr.) Goutam. (ed.) ()] *Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis. Applied Computational Intelligence and Soft Computing*, 12. 10.155/2018/8909357. 10. Ghosh, Monalisa & Sanyal, Prof(Dr.) Goutam. (ed.) 2018. 2018.

[Tate et al. ()] *Prediction of Dengue, Diabetes and Swine Flu Using Random Forest Classification Algorithm*, A Tate , U Gavhane , J Pawar , B Rajpurohit , G B Deshmukh . 2017.

[Rathi et al. ()] 'STUDY OF DENGUE INFECTION IN RURAL RAJASTHAN'. Manisha & Rathi , Masand , Alok Purohit . 10.14260/jemds/2015/993. `4.6849-6859.10.14260/jemds/2015/993` *Journal of Evolution of Medical and Dental Sciences* 2015.

[Aldallal and Al-Moosa ()] 'Using Data Mining Techniques to Predict Diabetes and Heart Diseases'. A S Aldallal , A A Al-Moosa . *4th International Conference on Frontiers of Signal Processing*, 2018. 2018. ICFSP. p. .

[World Health Organization (2014)] *World Health Organization*, `http://www.who.int/mediacentre/factsheets/fs117/en/` March 2014. 2014. 16 Oct 2019. Geneva. (Fact sheet no)