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

By Veena Kumari H M, Dr. Suresh D S & Dr. Dhananjaya P E

Abstract- 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, diarrhoea, cough, pain in the 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 a binary classification to classify dengue positive (DF) and dengue negative (NDF) cases using different ML techniques. The proposed work demonstrates that ensemble techniques of bagging, boosting, and stacking give 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 binary classification. The dataset is divided into 80% training and 20 % testing dataset. Performance parameters used for the analysis are accuracy, precision, recall, and f1 score, and compared the proposed model with other ML models.

Keywords: dengue fever, aedes aegypti, XGB, stacking, ROC, AUC.

GJCST-D Classification: DDC Code: 025.431 LCC Code: Z696

CLINICALDENGUEDATAANALYSISANDPREDICTIONUSINGMULTIPLECLASSIFIERSANENSEMBLETECHNIQUES

Strictly as per the compliance and regulations of:

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

Veena Kumari H M [α], Dr. Suresh D S [σ] & Dr. Dhananjaya P E [ρ]

Abstract- 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, diarrhoea, cough, pain in the 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 a binary classification to classify dengue positive (DF) and dengue negative (NDF) cases using different ML techniques. The proposed work demonstrates that ensemble techniques of bagging, boosting, and stacking give 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 binary classification. The dataset is divided into 80% training and 20 % testing dataset. Performance parameters used for the analysis are accuracy, precision, recall, and f1 score, and compared the proposed model with other ML models. The experimental results show that the accuracy of extended boost, random forest, and stacking is 98%, 99%, and 99% for the training dataset and 97%, 94%, and 98% testing dataset respectively. The extended metrics ROC, Precision-Recall curve and AUC better analysis

Keywords: dengue fever, aedes aegypti, XGB, stacking, ROC, AUC.

I. Introduction

Dengue 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.1

Fig. 1: A Female Aedes Aegypti Mosquito

Author α: Research Scholar, VTU, Belagavi Department of ECE Channabasaveshwara Institute of Technology, Gubbi, Tumkur, Karnataka, India.
e-mail: veenakumari.hm@cittumkur.org
Author σ: Professor Department of ECE Channabasaveshwara Institute of Technology, Gubbi, Tumkur, Karnataka, India.
e-mail: director@cittumkur.org
Author ρ: Professor, Department of General Medicine PES Institute of Medical Science and Research, Kuppam, Andhra Pradesh, India.
e-mail: Dhananjaya_pe@rediffmail.com

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].
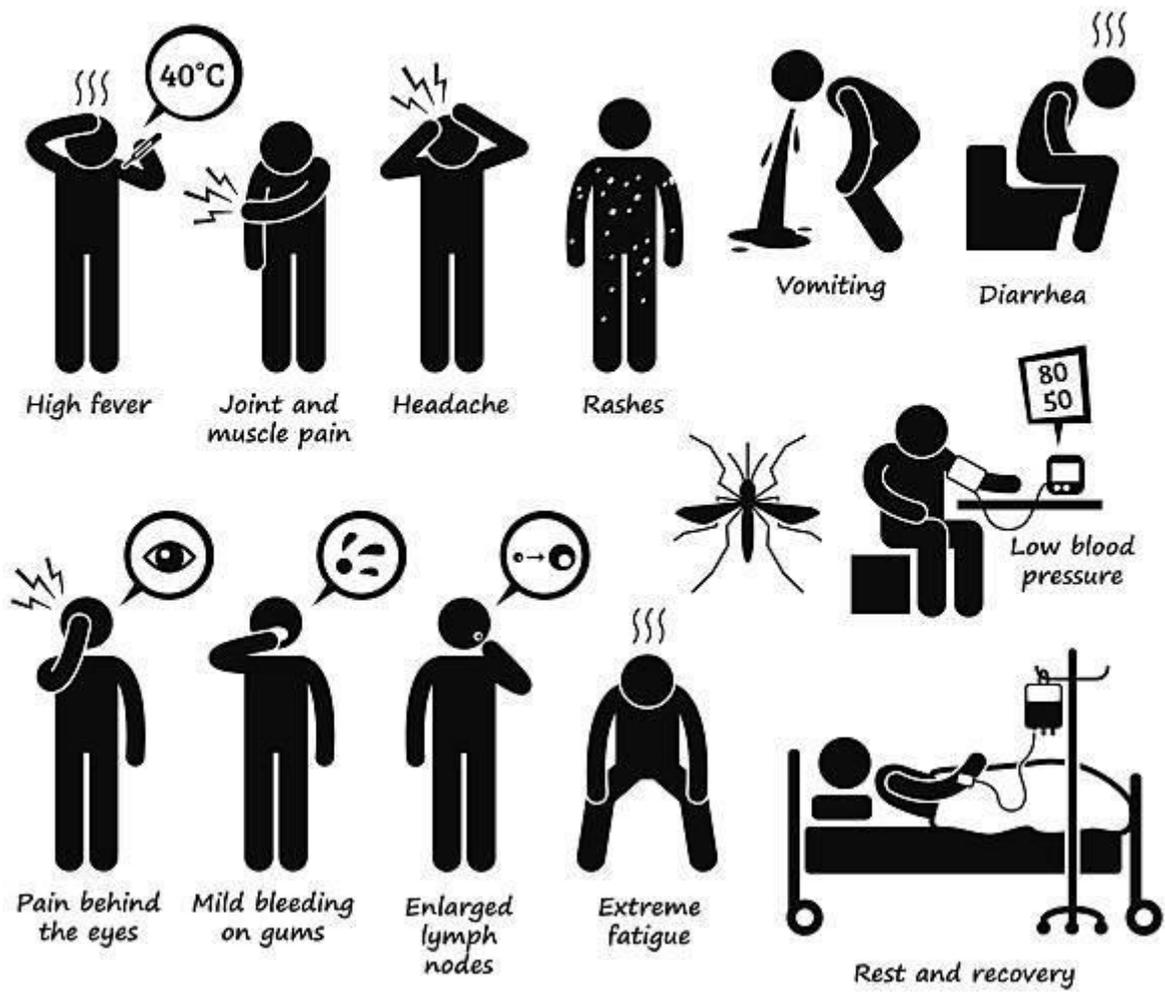

Fig. 2: Pictorial Representation of Dengue Fever Symptoms

According to the World Health Organization, Dengue fever is classified into four types: DENV1, DENV2, DENV3, and DENV4. The incubation period is 2 to 7 days [4]. The Dengue symptoms are high fever, joint and muscle pain, headache, vomiting, rashes, pain behind the eyes, diarrhea, etc. The dengue fever symptoms are shown in Fig.2.

Different ML algorithms are used for dengue fever classification such as NB classifier, K Nearest Neighbour, Decision Tree, Support Vector Machine, and Neural Networks. The proposed model demonstrates ensemble techniques called bagging, boosting, and stacking. The dengue binary classification is based on Extreme Gradient Boost (XGB), Random Forest by

majority voting, and stacking with different meta-classifiers. The techniques are analysed based on different performance measures called accuracy, precision, recall, F measure, and extended analysis done by the ROC curve, precision-recall curve, and AUC. The organization structure is as follows: Section II explains the work carried out and section III describes the proposed methodology. Performance analysis in section V and Section VI concludes the work.

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.

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.
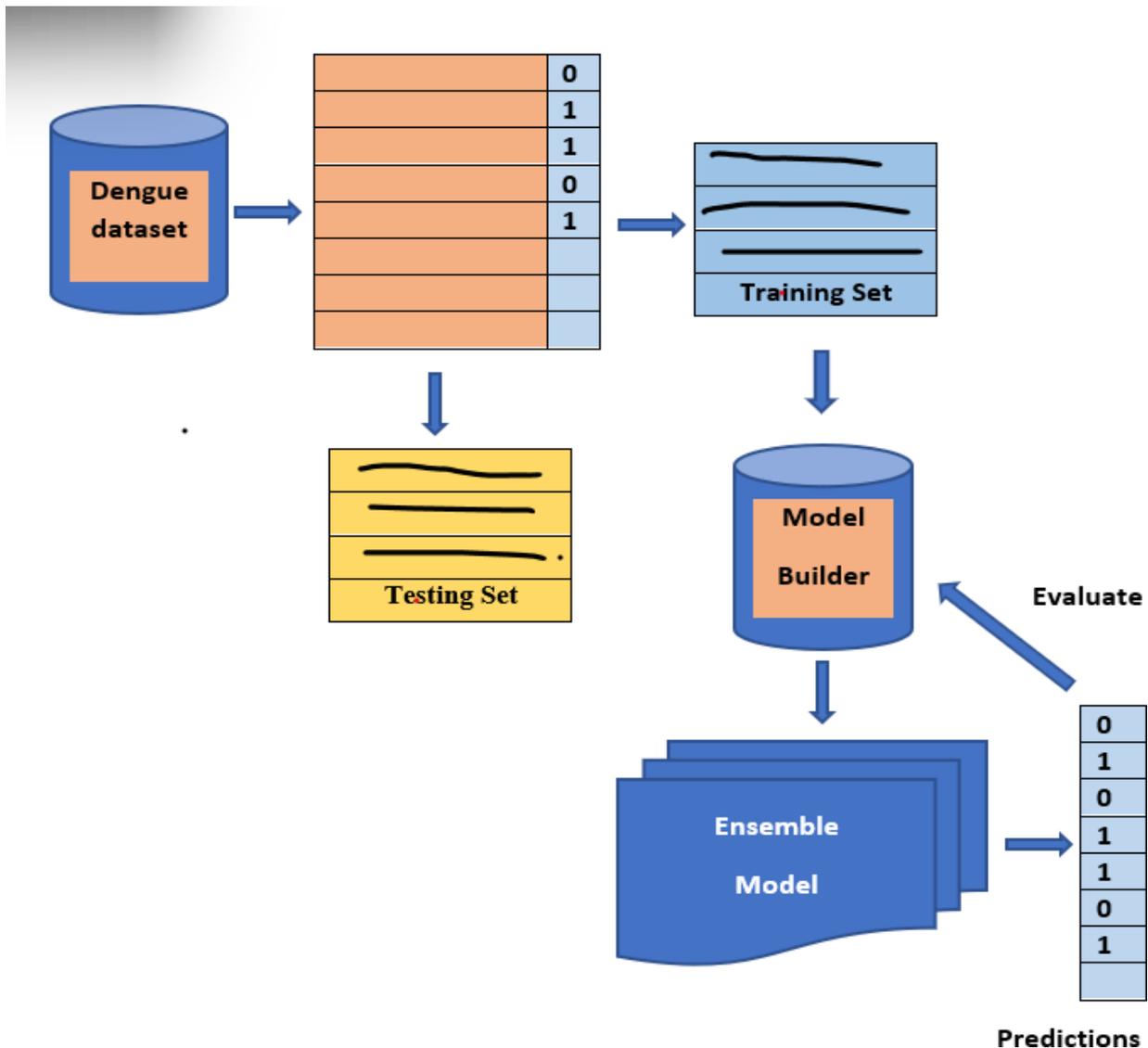

Fig. 3: An Ensemble Frame Work for the Prediction and Evaluation of Dengue Dataset

a) Data Acquisition and Analysis

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.

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.

Fig. 4: Diagnosis-Dengue Duo Card Test

It consists of 286 instances with 18 attributes and one target. The target consists of dengue patients and Non dengue patients. levels. The numerical value is assigned for each level like 0 for non- dengue patients (NDF), and 1 for Dengue patients (DF). The screenshot of the dataset is shown in Fig.5.

Fig. 5: The screenshot of the dataset

The target value consists of 140 cases of dengue infected and 146 non-dengue cases among 286 cases. The distribution is shown in Fig.6
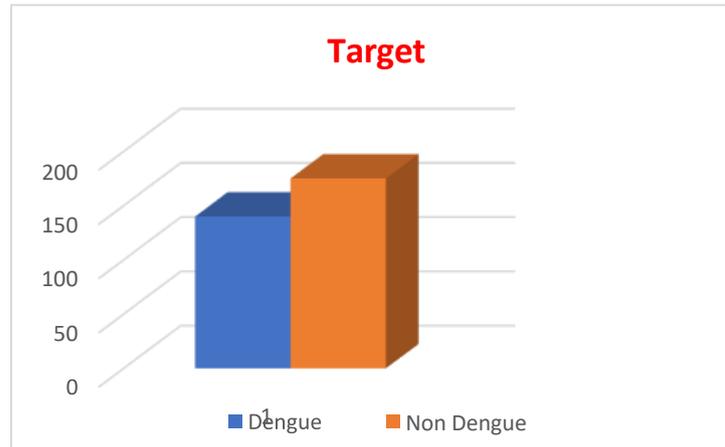
Fig. 6: Distribution of a Target Value

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.7. 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.

Table I: Major Clinical Features

Clinical Feature	No. of Patients
Fever	140
Headache	106
Myalgia	97
Arthralgia	94
Low Backache	83
Retro Orb Pain	71
Rashes	65
Vomiting	57
Pain Abdomen	41
Bleeding	39
Cough	30
Diarrhea	25
Sore Throat	16
Breathlessnes	6
Seizures	5

NO. OF PATIENTS

Fig. 7: Bar Chart Representation

b) Ensemble Methods

Ensemble means combining multiple models. This approach gives better performance compared to a single model. Thus, a set of models is used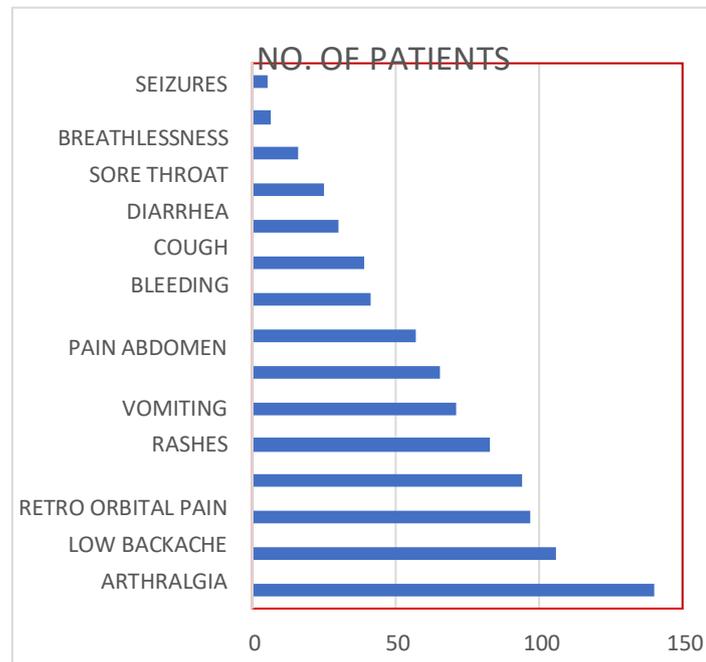 for predictions than a single model [7]. The main challenge is to obtain a base model which gives different kinds of errors. If the ensemble technique of bagging, boosting, and stacking are used for classification, high accuracies can be obtained. Bagging creates a different subset of training data from the sample training dataset & the final output depends on majority voting. e.g., Random Forest. Boosting the creation of sequential models by combining weak learners with strong learners and the finally constructed model has the highest accuracy e.g., XGBOOST and ADA BOOST

i. Random Forest Algorithm

Random forest is a supervised ML algorithm, used for both classification and regression. Random Forest is a bagging ensemble technique and each classifier in the ensemble model is a decision tree. RF constructs decision trees by a random selection of attributes at each node and then determines the split as shown in fig.8. Each tree votes and their majority vote are used for classification and the most popular class is returned. Random Forest Algorithm can handle the data set containing binary, continuous variables as well as categorical variables in case of regression and classification respectively. RF gives better results for classification problems. Random forest is a simple, fast, flexible, and robust model and it can handle missing values [10, 12].

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].
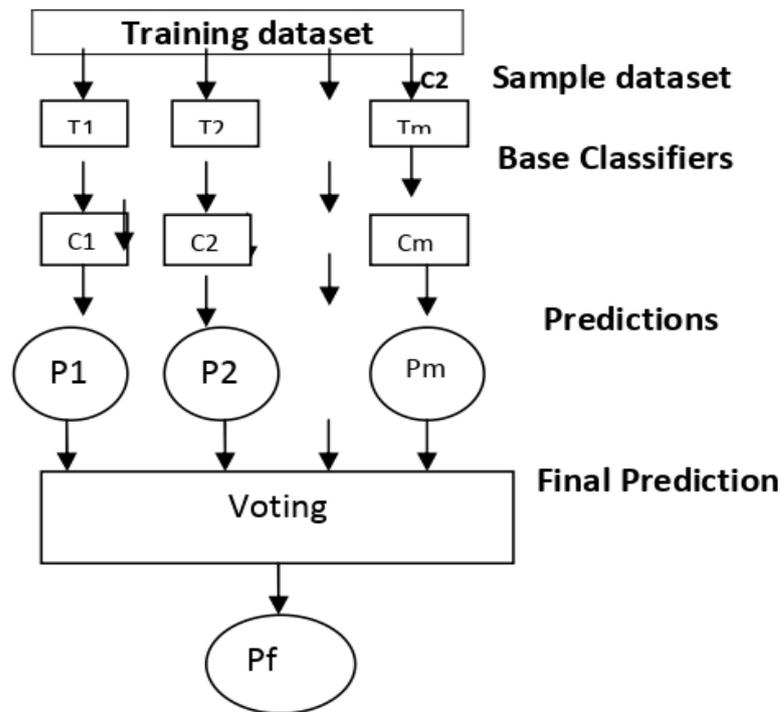
Fig. 8: Random Forest Algorithm Procedure

iii. *Stacking*

Stacking is an ensemble technique, which uses meta-classifiers to learn, the possible way to combine two or more base ML algorithms predictions. The base or level 0 classifiers consists of different ML algorithms and therefore stacking ensembles are generally heterogeneous classifiers. Level 1 classifiers are used as new features to train a meta classifier. An ensemble stacking procedure is illustrated in fig 9. The meta classifier can be any classifier [13]

Fig. 9: An Ensemble Stacking Procedure

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 fol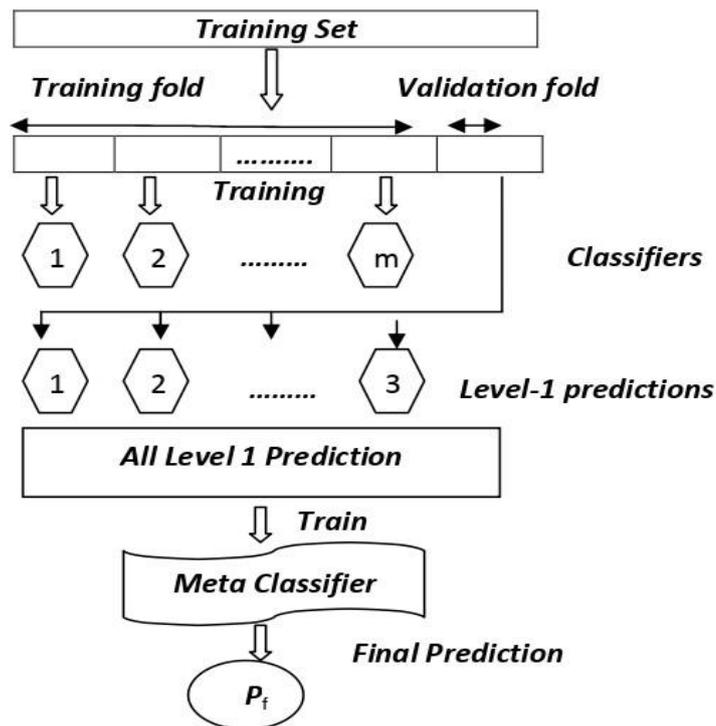ds are used to fit the level-1 classifier in k successive rounds. In every iteration, the level-1 classifiers are then applied to the remaining subset. The predictions of the base classifiers are then stacked and which is an input to the level-2 classifier.

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 (1), (2), (3), and (4), are used to calculate the metrics [17].

Table II: Confusion Matrix

		Actual	
		Dengue Infected	**Non-Dengue Infected**
Predicted	**Dengue Infected**	TP	FN
	Non-Dengue Infected	FP	TN

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (2)$$

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

$$F1\ Score- = \frac{2 * Precision * Recall}{Precision + Recall} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (4)$$

The confusion matrix and experimental score of the NB, SVM, KNN, XGB, RF, and Stacking models training dataset and testing dataset are shown in Fig.10.

Training Dataset	Testing dataset
NB	NB

Training Dataset — NB

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

Testing dataset — NB

Accuracy of Naivey bayes :
93.17

	precision	recall	f1-score
0	0.94	0.98	0.96
1	0.97	0.93	0.95

Training Dataset — KNN

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

Testing dataset — KNN

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

SVM

Accuracy of Support Vector Classifier: 97.5

	precision	recall	f1-score
0	0.96	0.98	0.97
1	0.99	1.00	1.00

SVM
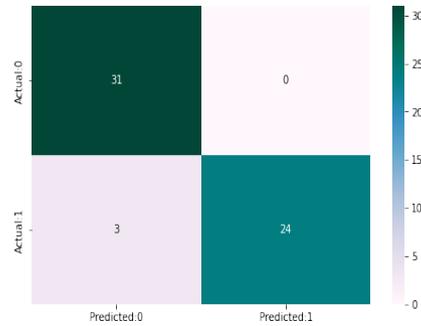
Accuracy of Support Vector machine: 89.65

	precision	recall	f1-score
0	0.91	0.98	0.94
1	0.98	0.89	0.93

XGB
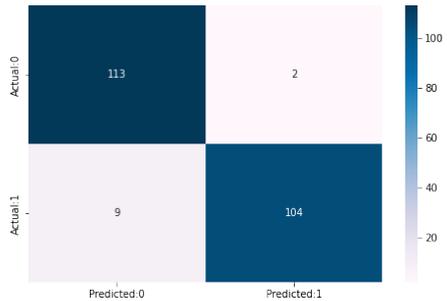
Accuracy of Extreme Gradient Boost :98.57

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

XGB
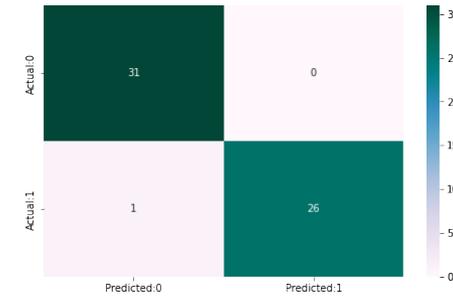
Accuracy of Extreme gradient Boost :97.80

	precision	recall	f1-score
0	0.97	0.98	0.97
1	0.98	0.96	0.97

RF

Accuracy of Random Forest: 99.12

```
        precision      recall    f1-score

0        0.98          1.00       0.99
1        1.00          0.98       0.99
```

RF

Accuracy of Random Forest : 94.82

```
        precision      recall    f1-score

0        0.97          0.98       0.97
1        0.98          0.96       0.97
```

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
```
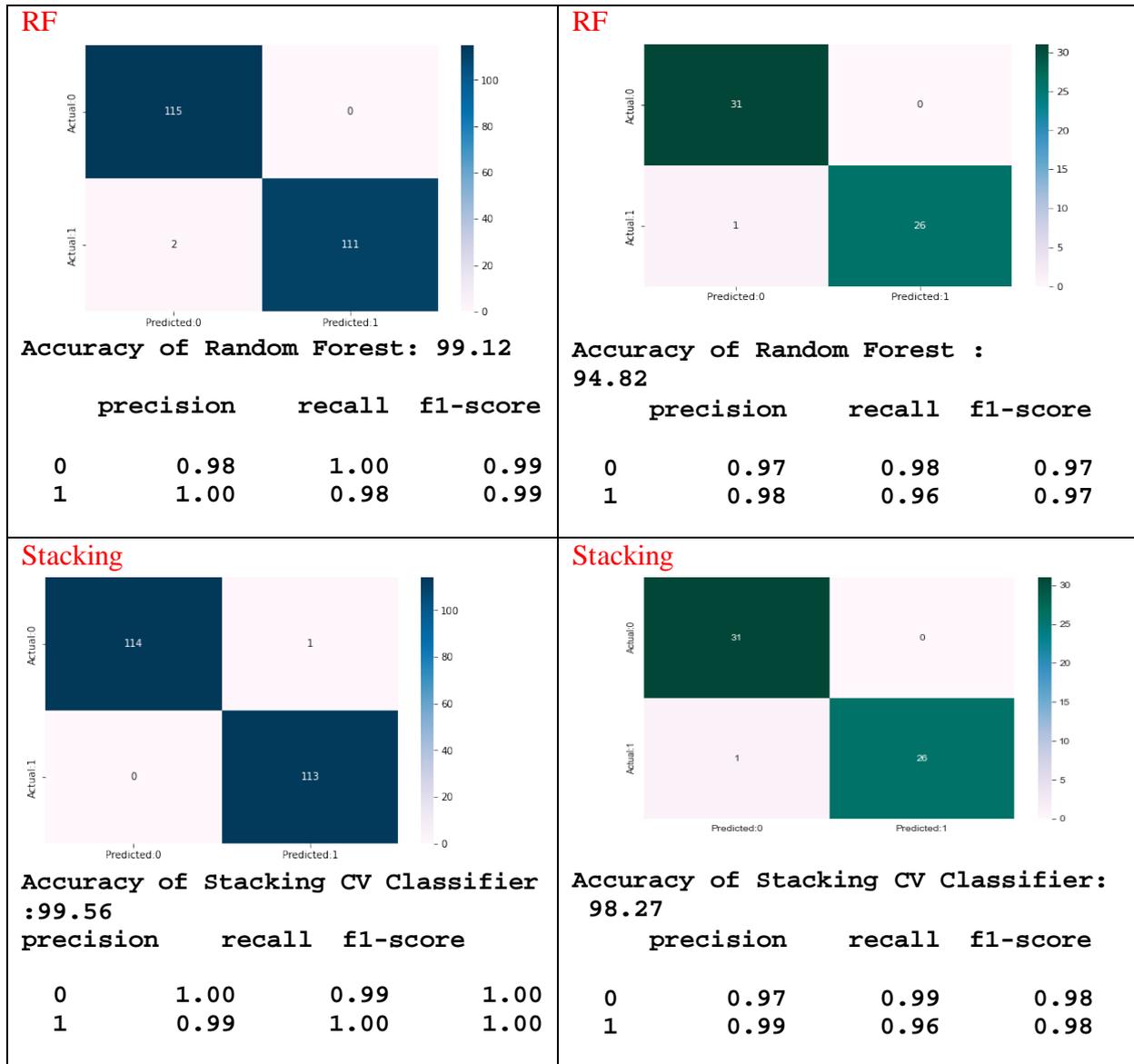
Fig. 10: Confusion Matrix and Experimental Results of Training and Testing Dataset of the Ensemble and Other M Models

Table III: Accuracy of Training and Testing Dataset

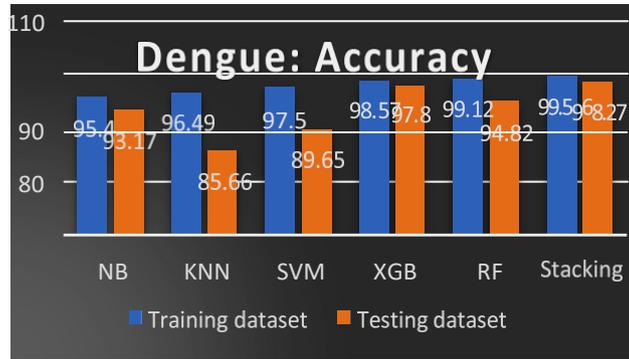| Classifiers	Training Dataset	Testing Dataset
NB	95.40	93.17
KNN	96.49	85.66
SVM	97.51	89.65
XGB	98.57	97.80
RF	99.12	94.82
Stacking	99.56	98.27

Fig.11: Accuracy Comparison of ML Models

The accuracy comparison of the training and testing dataset is shown in Table III and Fig. 11. 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.

Table IV: Precision, Recall and F1 Score of Training and Testing Dataset

Training dataset				
Classifiers		Precision (%)	Recall (%)	f1-score (%)
NB	NDF	93	98	95
	DF	98	92	95
KNN	NDF	93	100	97
	DF	100	93	96
SVM	NDF	96	98	97
	DF	99	100	100
RF	NDF	98	100	99
	DF	100	98	99
XGB	NDF	99	97	98
	DF	97	99	98
Ensemble Stacking	NDF	100	99	100
	DF	99	100	100

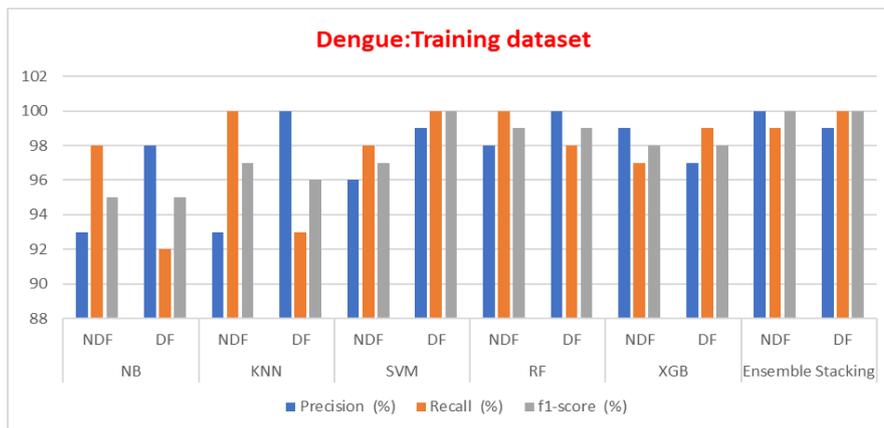| Testing Dataset				
Classifiers		Precision (%)	Recall (%)	f1-score (%)
NB	NDF	94	98	96
	DF	97	93	95
KNN	NDF	86	97	91
	DF	96	81	88
SVM	NDF	91	98	94
	DF	98	89	93
RF	NDF	97	98	97
	DF	98	96	97
XGB	NDF	97	98	97
	DF	98	96	97
Ensemble	NDF	97	99	98

Fig. 12: Training Dataset Precision, Recall and F1 Score Comparison of ML Models
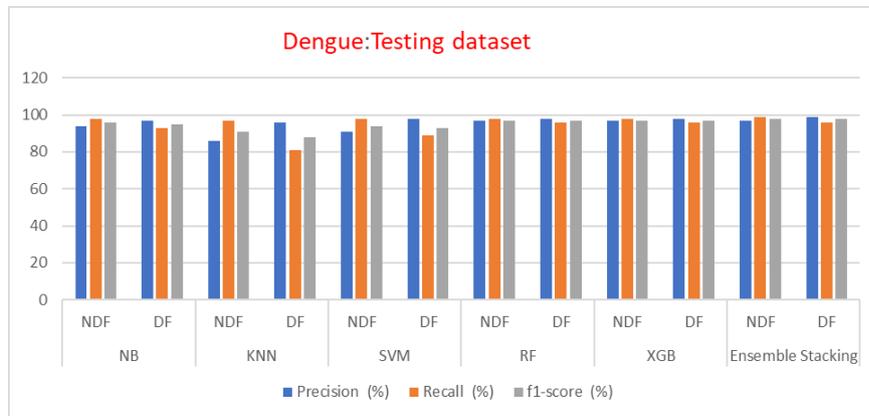
Fig. 13: Testing Dataset Precision, Recall and F1 Score Comparison of ML Models

The precision, recall, and f1 score for training and testing datasets are listed in Table IV and a comparison of an ensemble with other methods is shown in fig 12 and 13, which explains the ensemble methods give better performance for unseen data.

The Receiver Operating Characteristic curve and the Precision-Recall curve is a graphical representation of a, by calculating and plotting the false positive rate (FPR) Vs the true positive rate (TPR) and precision Vs recall for each classifier at various threshold values. The precision and recall curve for both training and testing datasets is shown in fig .14 and fig.15 correspondingly the ROC curve is shown in Fig 16 and Fig 17.
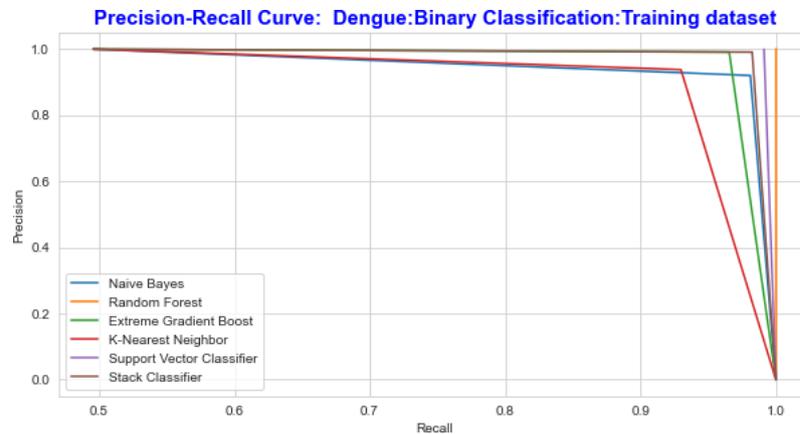

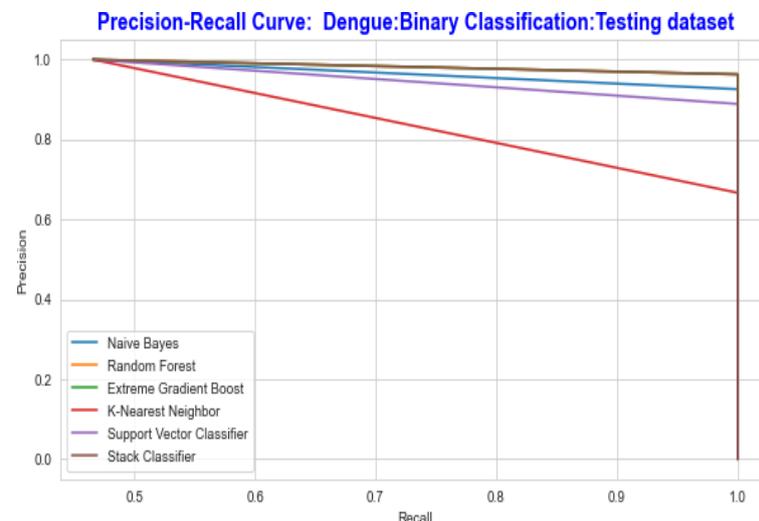Fig. 14: The Performance Comparison of the Training Dataset by Precision Recall Curve

Fig. 15: The Performance Comparison of the Testing Dataset by Precision Recall Curve
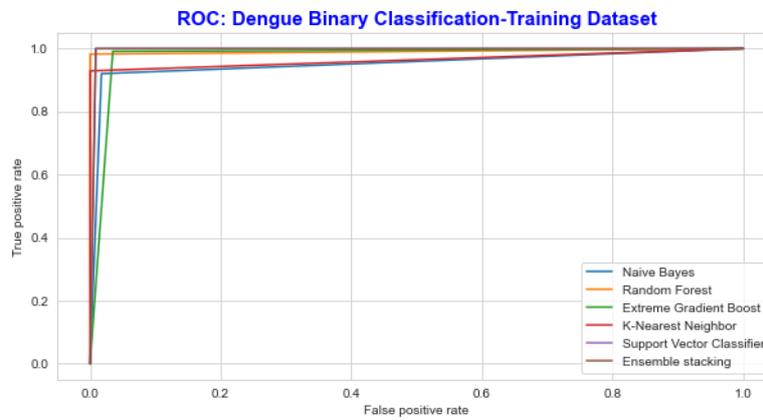
Fig. 16: The Performance Comparison of the Training Dataset by ROC Curve
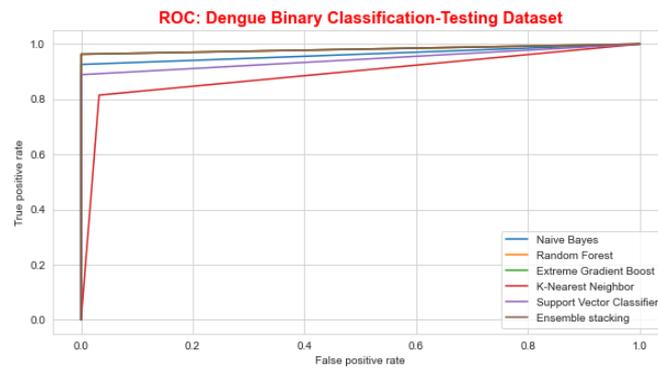

Fig. 17: The Performance Comparison of the testing Dataset by ROC Curve

The ability of the classifier can be measured by Area Under the (AUC) Curve. It is the summary of the ROC curve. High AUC indicates that the performance of the model is better, wherein differentiating between the positive and negative groups. AUC comparison with other classifiers is listed in TABLE 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.

Table V: Auc Comparision

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	0.9714	0.9781
Auc_Rf	0.9814	0.9914
Auc_Scv	0.9814	0.9868

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.

ACKNOWLEDGMENT

REFERENCES RÉFÉRENCES REFERENCIAS

1. World Health Organization. Dengue and severe dengue. Fact sheet no. 117, March 2014. Geneva: WHO, 2014. Available at: http://www.who.int/ media centre / factsheets/fs117/en/.[Accessed on 16 Oct 2019].

2. Yigzaw, Kassaye Yitbarek & Bellika, Johan. (2014). A communicable disease prediction benchmarking platform. 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI 2014. 564-568. 10.1109/BHI.2014.6864427.

3. Kalayanarooj S. Clinical Manifestations and Management of Dengue/DHF/DSS. Trop Med Health. 2011 Dec;39(4 Suppl):83-7. doi: 10.2149/tmh.2011-S10. Epub 2011 Dec 22. PMID: 22500140; PMCID: PMC3317599.

4. Rathi, Manisha & Masand, Rupesh & Purohit, Alok. (2015). STUDY OF DENGUE INFECTION IN RURAL RAJASTHAN. Journal of Evolution of Medical and Dental Sciences. 4. 6849-6859. 10.14260/jemds /2015/993

5. Aldallal, A.S., & Al-Moosa, A.A. (2018). Using Data Mining Techniques to Predict Diabetes and Heart Diseases. 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), 150-154.

6. Saba Bashir, Usman Qamar, Farhan Hassan Khan, IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework,Journal of Biomedical Informatics, Volume 59, 2016, Pages 185-200, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2015.12.0 01.

7. Agrawal, Kavita & Bhargav, G. & Spandana, E. (2021). Diabetes Diagnosis Prediction Using Ensemble Approach. 10.1007/978-981-15-5546-6_66.

8. Miao, Julia & Miao, Kathleen. (2018). Cardiotocographic Diagnosis of Fetal Health based on Multiclass Morphologic Pattern Predictions using Deep Learning Classification. International Journal of Advanced Computer Science and Applications. 9. 1-11. 10.14569/IJACSA.2018.090501.

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

10. Ghosh, Monalisa & Sanyal, Prof(Dr.) Goutam. (2018). Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis. Applied Computational Intelligence and Soft Computing. 2018. 1-12. 10.155/2018/8909357

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

12. Gupta, Gunjan & Adarsh, U & Reddy, N V Subba & Rao B, Ashwath. (2022). Comparison of various machine learning approaches uses in heart ailments prediction. Journal of Physics: Conference Series. 2161. 012010. 10.1088/1742-6596/2161/1/012010

13. Shaukat K, Masood N, Mehreen S, Azmeen U (2015) Dengue Fever Prediction: A Data Mining Problem. J Data Mining Genomics Proteomics 6: 181. doi:10.4172/2153-0602.1000181

14. Mesafint, Daniel & D H, Manjaiah. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications. 1-12. 10.1080/120 6212X.20 21.19746 63

15. Tate, A., Gavhane, U., Pawar, J., Rajpurohit, B., & Deshmukh, G.B. (2017). Prediction of Dengue, Diabetes and Swine Flu Using Random Forest Classification Algorithm

16. Nynalasetti, Konadala Kameswara Rao & Varma, Dr. G. P. Saradhi. (2014). A hybrid Algorithm for Dengue Disease Prediction with Multi Dimensional Data. International Journal of Advanced Research in Computer Science and Software Engineering. 14. 1033-1037.

17. H. M. Veena kumari and D.S. Suresh, "Comparative Study of Classification algorithms used for the Prediction of Non-communicable diseases," Int. J. Emerg. Trends Eng. Res., vol. 9, no. 7, pp. 917–920, 2021, doi: 10.30534/ijeter/2021/14972021.