Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765774

DENGUE PRONE AREA PREDICTION SYSTEM USING MACHINE LEARNING

Beulah Jayakumari R[1],MayaEapen[2], Vanitha R [3], Bhuvaneshwari G[4]Murugesan S[5], Merlin Vensiya V[6]

[1,2,3]Professor,[4,5]Assistant Professor, [6]Student

[1,4,5,6]Department of Information Technology, Tagore Engineering College, Chennai, India

[2,3]Department of Computer Science and Engineering, Jerusalem College of Engineering, Chennai, India

[1]hod.it@tagore-engg.ac.in,[6]meruvensi25@gmail.com

ABSTRACT

Machine learning (ML)is an emerging field in data science that predicts insights from a specific domain. Many tropical nations including Indiaare suffering from viral discrete diseases. One of the most threatening discrete diseasesis dengue virus. Hence it should be detected earlier for preventing the further spread around the prone area.In addition with exhausted pressure of covid-19 patients in hospitals early detection of dengue has become the real challenge to the physician due to similar symptoms of these severe viruses. Moreover currently existing algorithm failed to predict the disease due to lack of accuracy. The proposed method is to develop anefficient and accurate dengue prone area prediction classifier model.This is a multi-class classification model based on random forest algorithm which detects dengue prone area more effectively.The performance of the proposed system has been achieved using 12-folds cross-validation and split techniques. The performance metrics used in the proposed system includes precision, f-measures and accuracy. The results prove that the proposed system is more progressive by improvising in about 82.75 percent in accuracy, 85.7 percent in precision when compared with the existing system.

KEYWORDS: Random forest, Decision tree, Dengue diagnosis, Machine learning

I. INTRODUCTION

Machine Learning is a part of the Artificial Intelligence (AI) that gives the system the capability to study and improve automatically without being explicitly programmed. Machine learning focuses on the development of programs that can fetch data from real world, analyze it and gives appropriate solution by themselves. The backbone of such intelligent software is built using statistical learning techniques which is used to develop system intelligence.This paper deals about the detection dengue prone area using machine learning.

Dengue fever is one of the most common viral infectious disease caused by dengue virus (DENV) of the family flaviviridae, it is mosquito borne , having the aedes type mosquito as its carrier. It commonly occurs in tropical and subtropical areas of the world. The dengue disease is also known as break bone fever, because it causes severe bone and muscle pain but does not break any bone. The indications of the infection are headache, retro orbital pain, joint-pain, muscular pain and long lasting fever.

Dengue fever is categorized into two by the world health organization(WHO) , into types namely type1 and type 2 . Type 1 is the classical dengue fever and type 2 is the dengue hemorrhagic fever. Mild dengue fever causes a high fever and flue like symptoms. The severe form of dengue fever, also called dengue hemorrhagic fever can cause serious bleeding a sudden drop in blood pressure (shock) and death.
Dengue infection has endangered 2.5 billion populations all around the world. Every year there are 50 million people who suffer from it globally. Pakistan has been victim of this rapidly growing sickness from last few years , large number of cases was marked especially in Lahore in 1994 at Karachi Pakistan's first case of dengue was appeared.

The global incidence of dengue has grown dramatically with about half of the world's population now at risk .although an estimated 100-400 million infections occur each year, over 80% are generally mild and asymptomatic. Dengue has distinct epidemiological patterns, associated with the four serotypes of the virus. These can co circulate within a region and indeed many countries are hyper-endemic for all four serotypes. Dengue has an alarming impact on both human health and the global national economics. Recovery from infection is believed to provide lifelong immunity against that serotype. However, cross-immunity to the other serotypes after recovery is only partial, and temporary. Subsequent infections by other serotypes increase the risk of developing severe dengue.

The starting Symptoms of DHF is fever which long last for more than 2-7 days along with the signs of like revealing of plasma, shock and weak pulse. In the earlier days it was more difficult to found the difference between the typ1 and the type 2 fever. But fortunately now emerging of ML techniques have been implemented in the early detection of dengue Diagnosis naïveb ayes (NB) classifier, k-nearest neighbor (KNN) technique, multilayered technique and support vector machine (SVM). In this proposed work we have implemented randomforest classifier based dengue diagnosisand the results evaluation obtained are based on the measures called accuracy, precision, f-measures.

The main contribution of the proposed work includes

- An efficient feature selection method named Recursive feature elimination (RFE) is presented to determine most promising features.
- The ensemble based random forest classifier model is built for determining dengue prone areas.
- The performance evaluation is done through 12-folds cross-validation and split techniques.
- The performance metrics used in the proposed system includes precision, f-measures and accuracy. The experimental results established superiority of the proposed random forest classifier model.

II. RELATED WORK

Sugandhi C, et al.[1] deals with an analysis of a population of cataract patient database. This system make use of the linear discriminate model. It was found to be most reliable trained

model for the 30 dengue fever data with a true positive weight of 83.3 percentage and false negative of 16.7 percentage and accuracy validation was about 50 Percentage.

Yasodha P and Kannan M, [2] analyse diabetic patient database using weka tool. It describes the development of non-linear autoregressive moving average with exogenous input (NARMAX) models in diagnosing dengue infection. This analysis show that the NARMAX model yield least accuracy compared to the autoregressive moving average with exogenous input (AIRMAX) model in diagnosis intelligent system.

David S. K. et al.[3]Solanki et al [4] done a comparative analysis of data mining tools and classification technique for dengue diagnosis and developed an intelligent expert system for dengue fever prediction based on symptomatic features to detect dengue before pathological test.

K. Lee, et al.[5] talk about real-time disease surveillance using Twitter data for demonstration on flu and cancer. It develop an inexpensive, but robust and user friendly diagnostic device that can be used for the detection of dengue fever at a molecular level. The result indicated that this paper diagnostic device was capable of detecting the RT:LAMP products in the buffer system with the concentration of three hundred ng/ml.

Xue H et al.[6] study based on regional level influenza in Twitter and Machine Learning method. It was implemented using random forest with grid search algorithm to diagnose the class or level of dengue fever by tuning the hyper parameters with the grid search approach for the prediction. Comparative analysis shows that the ensemble model gives least precision and recall and F1 score for the multi label classification of dengue than other ML models.

Wakamiya et al.[7] proposed methodology for tweet classification toward twitter-based disease surveillance. The dengue fever and dengue hemorrhagic fever clinical symptoms were studied from an expert system perspective, and represented in a tree like structure amenable to expert system application.

A less a and Faezipour [8] describes preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports. This study was conducted with the primary objective to test and evaluate 8 different classification algorithms used to predict dengue virus infection cases into three classes, dengue fever, dengue hemorrhagic fever and dengue shock syndrome. The result of the comparison shows that the neural network algorithm has the best accuracy that was over other algorithms.

Vijayarani et al [9]Dhamodharan S et al.[10] Joshi [11]proposed an liver disease prediction using bayesian classification. This study uses a fuzzy logic approach for the patient to get further notification if they are suspected with the liver disease. This is considered to be a reliable system for liver diseases.

Rigauet al [12] Tanner et al. [13] predict the diagnosis and outcome of dengue fever in the early phase of illness using decision tree approach. This classify 1200 patients with 6 remarkable features and achieved an accuracy of 84 percent.

Durairaj M, et al.[14] study about data mining applications in healthcare sector . The main goal of this research work is to predict people who are affected by dengue depending upon categorization of age group using K- means clustering algorithms. The dataset were substituted along with missing values for nominal and numeric attributes with mode and mean values.

Kumar M. N., et al.[15] construct an alternating decision trees for early diagnosis of dengue fever. This computational intelligence based methodology predicts the diagnosis in real time, minimizing number of false positives and false negatives. The predictive models developed using this methodology has found to be less accurate than the state of the art methodology used in the diagnosis of dengue fever.

Farooj W et al. [16] used data mining techniques for the efficient classification of the dengue fever type on the basis of weighted features , minimum cost and source availability. This produces the accuracy of 99.4 percent. The drawback of the experiment is more type 2 error.

Shaukat et al [17]T. F. M. de Lima et al.[18], used tool for the modeling and simulation of dengue spatiotemporal dynamics. This adopts non-invasive method, the generalized regression neural network (GRNN) for training the existing patient databasein order to perform diagnosis and prognosis of stages. It is observed that KNN provides 63 percentage of accuracy GRNN provides 87percentageaccuracy.

Gislason PO et al.[19] deals with random forest classification of multisource remote sensing and geographic data. This research aims to develop a prediction framework that supports early diagnosis of diseases.

They collected distinct attributes are used for clustering the model. They implemented the technique decision tree and nearest neighbor models, but the limitation of this proposed work is low scalability.

III. PROPOSED WORK

Dengue is a life threatening disease prevalent in several developed as well as developing countries like India. Dengue is a virus born disease caused by breeding of Aedesmosquito. The main objective of the proposed work is to develop an efficient dengue prone area prediction system classifier.

The proposed architecture consists of four stages which include data acquisition, data pre-processing, data analysis, data prediction and data visualization is shown in figure 1: System architecture.

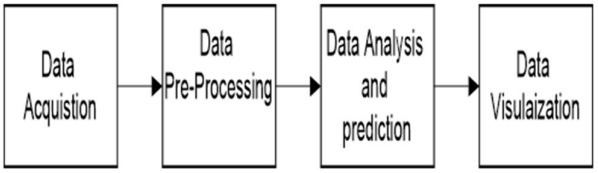

Figure 1 System architecture

For this analysis dataset are acquired from Kaggle(https://www.kaggle.com/ /kaggle/input/dengueai/dengue_features_train.csv, /kaggle/input/dengueai/dengue_labels_train.csv)

The dataset comprises of 24 features with 1457 records.
Data acquisition stage is followed by preproessing stage, transforms the raw data into clean data set for analysis by machine learning model. The two main pre-processing approaches used in the proposed work are:

i. Finding the missing data
ii. Splitting the entire data into the train and test

The third stage in data analysis is feature selection, here the data set comprises of 24 features, among only 12 highly correlated features are extractedfor the construction of accurate classifier model. It uses an iterative approach called wrapper type Recursive Feature Elimination (RFE)using cross validation automatically extracted the number of features based on best mean score.RFE works by searching for a subset of features with all features in the training dataset and successfully removes features until the desired count reaches. Eliminates less i feature based on weighted feature importance (WFI).

The first case is initiated with pair wise comparisons to rank the importance of features, and then distance-based inconsistency reduction was used to refine the weights assessment and make comparisons more precise. In the next step, calculate the weights through the fully-consistent or almost consistent pair wise comparison tables.For the second case, a novel concept of feature domain over lapping has been introduced. It can measure the feature discrimination power. This model is based on the assumption that less overlapping means more discrimination ability, and produces weights characterizing the importance of particular features. For both cases Weighted Support Vector Machines are used to classify the data. Both methods have been tested using two benchmark data sets, Iris and Vertebal. The results were especially superior to these obtained without weights. The 12 highly correlated features used are dengue precipitation, name of the city, year, week start date, week end date , maximum temperature, minimum temperature , average temperature, southwest city centroid , north east city centroid , south east city centroid and north west city centroid.

Dengue Precipitations wererepresented in mm (millimeter), maximum temperature, minimum temperature and average temperature were represented in c(Celsius) The notation used for the selected features are depicted in the table 1.1

Final stage in data analysis is statistics prediction. It is executed using ensemble based classifier referred to as Random forest. This supervised classifier model combines several decision trees to predict the final output

Table 1: Selected Feature table

S.no	Name of the feature	Notation
1	dengue precipitation,	station_precip
2	name of the city,	city
3	year,	year
4	week start date,	week_start_date
5	week end date ,	week_end_date
6	maximum temperature,	station_max_temp
7	minimum temperature ,	station_min_temp
8	average temperature	station_avg_temp
9	southwest city centroid	ndvi_sw
10	north east city centroid	ndvi_ne
11	south east city centroid	ndvi_se
12	north west city centroid	ndvi_nw

In widespread, decision tree classifier is based on bagging or bootstrap aggregating approach which takes into account best random subset of statistics. But random forest classifier uses both subset and function set of facts randomly with substitute. This overcomes bias and variance in single and complex tree respectively. As a result ensemble primarily based classifiers are desired.

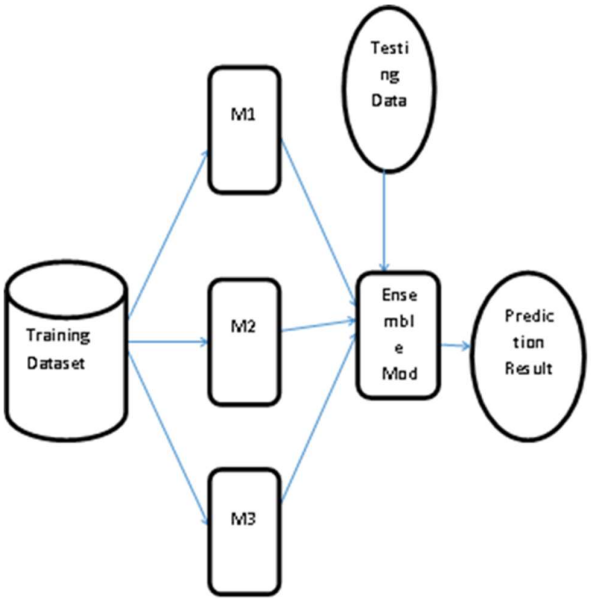

Figure 2 Random forest model for dengue prediction

It works in two phases namely training and testing phase. During training phase, the training dataset (70%) D is split into 'n' training subsets D1, D2,…….Dn using a probabilistic approximation technique called the Reservoir sampling method. Then each training subset produces a base classifier. During testing phase, remaining 30% of dataset are used to evaluate the proposed ensemble learning method. It combines the result of base classifiers to predict the final output. The predicted final output is determined by major dengue prone areas among the weighted output of base classifiers. The algorithms used in dengue prediction are shown in figure 3 and figure 4.

GenDecTree(sample S, feature F)
STEP:
1.**If**
stopping_condition(S,F)=true
then
a. leaf = create Node( )
b. leaf label = classify(s)
c. return leaf
2. root =create Node( )
3. root.test_single_ node_ tree = find best
split(S,F)
4. V={v\V a possible outcome
cfroot.test_single_node_tree}
5. **for** each value vEV :
a. S ={s\root.test_single_node_tree(s)=v
and Ses};
b. child = Tree growth(S ,F);
c.add child as descent of root and label
the edge {root-child}as v
6.**return** root

Figure 3 Decision tree algorithms for dengue prediction

A training set S =(x1,y1)……..(xn,yn) features F, and number of trees in forest B
Function RANDOM FOREST(S,F)
H-0
For I E 1……B
Do
Si-a bootstrap sample from S
Hi-RANDOMIZED TREE LEARN(Si,F)
H-HU{hi}
End for
Return H
End function
Function RANDOMIZED TREE LEARN(Si,F)
At each node :
f-very small subset of F
return the learned tree
end function .

Figure 4Random forestalgorithm for dengue prediction

IV. PERFORMANCE RESULT

Figure 1 shows the comparison of total number of dengue cases recorded intwo cities namely San Juan and Iquitos are 936 and 520 respectively.

```
San Juan:  936
Iquitos :  520
```

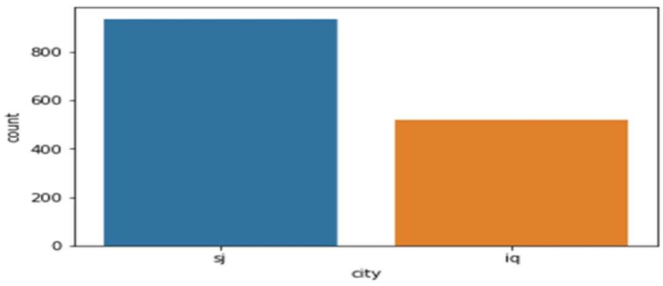

Figure 1.City VsTotal number of dengue cases

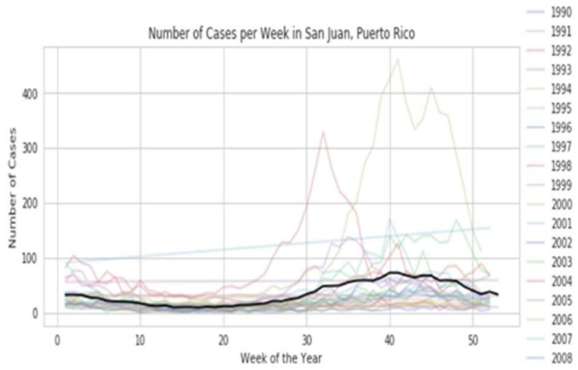

Figure 2: Week of the year Vs Number of dengue cases in cities

Figure 2 represents the increasing number of dengue cases that has been diagnosed during every week of the year in cities San Juan and Puerto Rico

This plotting illustrates that in the year 1992 the number of dengue cases reached the peak level diametrically 2007 is the year with the least number of dengue cases and 2005 is the moderate year of total number of dengue cases.

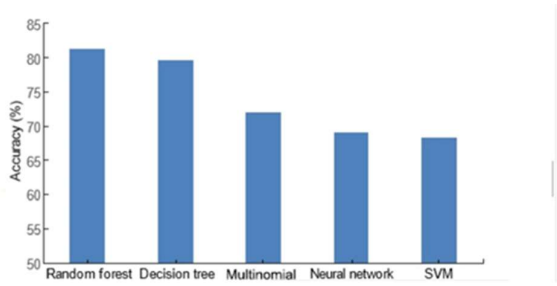

Figure 3: Various Classifier model Vs% of accuracy

Figure 3 shows the comparison between various classifier models in terms of accuracy. The results proves that multi class random forest based classifier gives more accuracy 85% when compared with decision tree(82.7%), multi nominal, neural network(63.8%) and support vector machine(SVM)(67.8%) like other conventional classifier models.

Figure 4: Number of folds Vs Accuracy

In figure 4 we made a comparison for number of folds verses accuracy for measuring the consistency of the random forest model, which provided nearly linear performance to the prediction of all thecases.

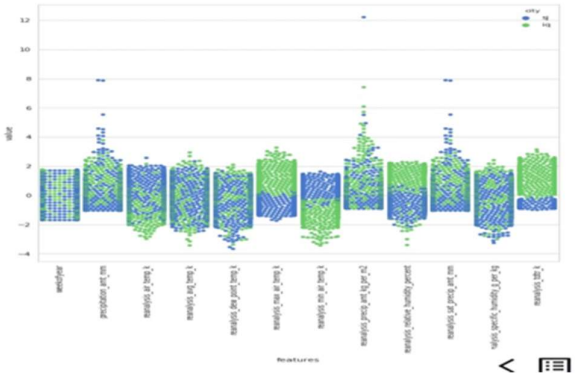

Figure 5: Features Vs Value

Figure 5 represents plotting of all the features considered in the dataset with their values.

V. CONCLUSION

The developed dengue prone area prediction system classifier model is more efficientin terms of precision, f-measures and accuracy. When compared with the traditional classifier the proposed ensemble based classifier out performs well during validation process .the performance evaluation is carried out using k-cross validation technique. The performance metrics used for the evaluation areprecision, f-measure, accuracy. The results prove that the proposed system out performs in about 85.7 percent in precision, 82.75 percent in accuracy when compared with the existing system.

In the future, we extend the prediction analysis to heterogeneous ensemble base classifier model. Also apply this classifier model to any other regions for finding the preventive and remedial measure in outbreak of diseases in near future with the attributes as symptoms of viral dengue fever.

REFERENCES

1. Sugandhi C , Yasodha P , Kannan M , Analysis of a Population of Cataract Patient Database in WEKA Tool , International Journal of Scientific and Engineering Research ,2(10) ,October ,2011.
2. Yasodha P, Kannan M, Analysis of Population of Diabetic Patient Database in WEKA Tool,International Journal of Science and Engineering Research, 2 (5), May 2011.
3. David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.
4. SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014
5. K. Lee, A. Agrawal and A. Choudhary, "Real-time disease surveillance using Twitter data: Demonstration on flu and cancer", Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), pp. 1474-1477, 2013.
6. Xue H, Bai Y, Hu H and Liang H, "Regional level influenza study based on Twitter and machine learning method", PLoS ONE, vol. 14, no. 4, pp. 231-253, 2019
7. S. Wakamiya, M. Morita, Y. Kano, T. Ohkuma and E. Aramaki, "Tweet classification toward Twitter-based disease surveillance: New data methods and evaluations", J. Med. Internet Res., vol. 21, no. 2, Feb. 2019
8. Alessa and M. Faezipour, "Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports?: Prediction framework study", JMIR Public Heal. Surveill., vol. 5, no. 2, pp. 1-17, 2019
9. Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, 1(3): 735-741, 2013.

10. Dhamodharan S , Liver Disease Prediction Using Bayesian Classification , Special Issues , 4th National Conference on Advance Computing , Application Technologies, May 2014
11. Joshi J, Rinal D, Patel J, Diagnosis And Prognosis of Breast Cancer Using Classification Rules, International Journal of Engineering Research and General Science,2(6):315-323,October 2014
12. Rigau-Pérez J G, "Dengue and dengue haemorrhagic fever" , The Lancet 19: 971-977, 1998
13. Tanner L, Schreiber M, Low JG, Ong A and Tolfvenstam T "Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness. PLoS Neglected Tropical Disease 12, 2008.
14. Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.
15. Kumar M. N., Alternating Decision trees for early diagnosis of dengue fever. Ar Xiv preprint arXiv:1305.7331,2013
16. Farooqi W, Ali S and Abdul W Classification of Dengue Fever Using Decision Tree. VAWKUM Transaction on Computer Sciences 3: 15-22, 2014.
17. Shaukat K, Masood N, Mehreen S and Azmeen U "Dengue Fever Prediction: A Data Mining Problem. J Data Mining Genomics Proteomics 6: 181. doi:10.4172/2153-0602.1000, 2015
18. T. F. M. de Lima, R. M. Lana, T. G. De SennaCarneiro, C. T. Codeço, G. S. Machado, L. S. Ferreira, et al., "DengueME: A tool for the modeling and simulation of dengue spatiotemporal dynamics", Int. J. Environ. Res. Public Health, vol. 13, no. 9, pp. 1-21, 2016
19. Gislason PO, Benediktsson JA, Sveinsson JR (2004) Random forest classification of multisource remote sensing and geographic data. Geoscience and Remote Sensing Symposium 2004 IGARSS'04 Proceedings 2004 IEEE International Vol 2.