Evaluation of Machine Learning Approaches for Prediction of Dengue Fever

Chapter · October 2022

DOI: 10.1007/978-981-19-3035-5_13

2 authors, including:

Tasmiah Rahman
Daffodil International University
4 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Evaluation of Machine Learning Approaches for Prediction of Dengue Fever

Tasmiah Rahman[1], Md. Mahmudur Rahman[2]
Department of Computer Science and Engineering, Daffodil International University,
Dhaka,Bangladesh
[1]tasmiah15-8805@diu.edu.bd,[2]mahmudur3720@diu.edu.bd

Abstract.Dengue is a mosquito-borne, deadly viral disease that is a major threat to public health all over the world. Dengue and Covid -19 symptoms are almost same and sometimes people are confused about which disease they are infected with. This year in Bangladesh dengue and covid-19 patients have been increasing at an alarming rate and most of the time people didn't properly recognize the disease. A developing country like Bangladesh has faced many difficulties to handle this situation. The target of this research work is to analyze the symptoms and predict the chances to get infected with dengue fever. Machine learning techniques are widely utilized in the health industry to detect fraud in treatment at lower cost, predictive analysis, cure the disease. Four machine learning algorithms are used which are Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest to predict dengue fever based on symptoms. The results were compared for percentage split and K-fold cross-validation method for before and after applying principal component analysis. The experimental result shows that the Support vector machine algorithm provides the highest performance compared to others algorithms.

Keywords: Bangladesh, machine Learning, support vector machine, principal component analysis, dengue, cross-validation, percentage split.

1 Introduction

Dengue is a mosquito-tainted viral illness that can spread rapidly under tropical climatic conditions. Primarily imparted effectively through the bite of female mosquitoes named 'Aedesaegypti', dengue increasing predominance is principally because of varieties in temperature, rainfall, and unplanned rapid urbanization. Dengue epidemic first clinically identify in Asia, Africa and North America occurred nearly concurrently in 1780's condition as "Break Bone Fever". 3.9 billion people all over the world are in danger of infection with dengue viruses[1]. Each year almost 50 million people are suffering from dengue fever worldwide. Today, dengue cases have multiplied all throughout the planet; regardless, the genuine numeral of dengue occurrence are either never or some of the time erroneously dispersed. Dengue virus is found in torrid and sub-torrid weather around the globe,

especially in townish and semi-townish areas. Outrageous dengue is a primary wellspring of passing among kids and authentic sickness in some Latin American and Asian nations. Dengue fever has no specific treatment but early prediction can reduce the damage of health. Dengue is a mosquito-borne disease which is specially found in sub-tropical and tropical areas in the whole world.

This year dengue fever turned into a terribly situation in Bangladesh. The directorate general of health services (DGHS) announced that 6,450 patients hospitalized this year in the country. 28 people deaths among them [2]. Most of the people in the country are affected in DEN-3, that is a dengue virus variant which can rise the mortality risk. Bangladesh outbreaks of DEN-1 and DEN-2 variants, previous year but this year DEN-3 is more prevalent. However, due to the Covid-19 epidemic this year, dengue has become riskier and more complex because dengue fever patients show the same symptoms like Covid-19 pandemic, and it makes difficulties for diagnosis this disease [3].

Machine learning is one the important and motivating extent of research work with the legitimate of invention comprehensive knowledge from huge number of data sets. For identify hidden knowledge in health data, healthcare is most popular field application in machine learning. Machine learning is utilized in health industry to detect fraud in treatment at lower cost, health insurance, cure the disease. In this work, different machine learning models are utilized to acquisition the best model which predict dengue fever more accurately.

2 Related Work

Significant amount of deed was done by numerous researches all over the world on disease prediction utilizing machine learning and data mining tools. Mostly, they have used one of the two, either data mining or machine learning, to create a model, often times by utilizing several techniques of their choice. The following is a brief description of some of those works.

Permatasari et al. in their research wok they utilized Quadratic Discriminant Analysis and mapping the occurrence within 5 areas in Palembang founded on different attributes for example Gender, Blood group, Age etc.66.7% was the accurate percentage of the mapping outcome [4]. Saif et al. published an article where they utilized 1047 data. They applied Support Vector Machine and Random Forest Classifier and found the result is almost same which are 0.694 for Support Vector Machine and 0.681 for Random Forest Classifier [5]. Ramandeep et al. applied Clustering in their research work which is a part of datamining. They classified patients information by gender, age between 25-60 and attribute. In this work they find the maximum ratio of male gender. Dengue fever in most case rely

on Rashes and Myalgia [6]. Kamran et al. did a work where they collected the dataset from DHQ hospital Jhelum, Pakistan. They utilized Naïve Bayesian, J48, Random Tree and SMO algorithms and find the highest accuracy 88% from J48 and 92% from Naïve Bayesian Algorithms [7]. Fatimah et al. developed a dengue fever prediction system by utilizing clinical symptoms. They utilized Multilayer feed-forward neural network and find 90% accuracy of prediction [8]. Jorge et al. compared two machine learning algorithms which are ANN and SVM that helps in medical diagnosis. ANN-MLP produced the best performance and achieved 96% accuracy,97% specificity and 96% sensitivity [9]. Abdiel et al. proposed a model to predict dengue fever outbreak where they used ANN algorithm that was trained with genetic algorithm. They discuss that ANN algorithm had an excellent predictive ability[10]. In another work Abdul et al. utilized Discriminant Analysis (DA) and Principal Component Analysis (PCA) for predict dengue possibility. DA method found the accuracy between 93-98%,Specificity between 94-100% and Sensitivity between 75-89% [11].Sanjudevi et al. used Decision Tree, Support Vector Machine and SMO for prediction of dengue fever in their research work. They achieved the highest accuracy from Support Vector Machine which is 99% [12] . In another research work permatasari et al. was used two different machine learning techniques that are Artificial Neural Network (ANN) and Random Forest (RF) with 10 -Flod cross validation method. They found the highest accuracy 58% from Random Forest Classifier [13] . Shalini et al applied Decision Tree, Artificial Neural Network and Naïve Bayes algorithms in their research work where the used 110 patient information for prediction of dengue fever. They find the highest accuracy from ANN algorithm which is 79.09% [14] . A.Shameem et al. utilized Support Vector Machine for analysis dengue fever. They found 0.9042 accuracy in this research work [15].Akey Sungheeta et al.in their research work for detection and classification of diabetes rethinopathey they utilized Image Processing,SVM,Naive Bayes and proposed deep features CNN. They found 97% accuracy from proposed deep features CNN which is the highest among all methods[16].In another research work Milan Tripathi et al. Utilized three machine learning methods which are Naive Bayes,SVM and LSTM for detecting sentiment of covid-19 tweets analysis.where they utilized 4035 data.After analysis they found the highest accuracy from LSTM classifier which is 79%[17].

3 Proposed Method

In this study, a several number of machine learning techniques are analyzed for dengue fever prediction. The first quest was to acquire a dataset with numerical value. Then prepare the dataset for utilize. At the end, using percentage split,10-fold cross validation, performance matrix and various classifications are evaluated for finding the best model which can predict dengue fever more accurately. Fig 1. shows the work flow of our system.

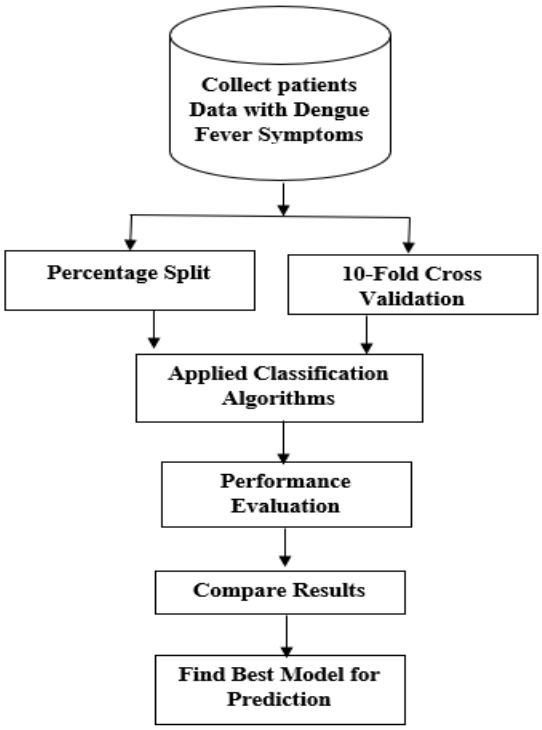

Fig. 1. Workflow of the proposed system

3.1 Dataset

Dataset is called the heart of any research work. This is one of the most important part. This dataset is collected by using survey and internet from the people who are affected in dengue fever and who have dengue fever symptoms. Dengue fever symptoms are used for making this dataset. By surveying we are collected 1116 people's information where 646 are not affected in dengue fever and 470 people are affected in dengue fever which is shown in Fig.1. Here, "0" means not affected by dengue fever and "1" means affected by dengue fever.Symptoms of dengue fever which we used to make our dataset is given in Table 1.

Table 1. Symptoms of Dengue Fever

Most Common	Moderate	Severe
Fever,Cough,Fatigue,Muscle Pain	Sore Throat,Rashes,Severe Headache, Loss of Taste,Nausea,Pain back of the eyes	Bleeding,Low Heart Rate, Platelet Decrease,Black Closet

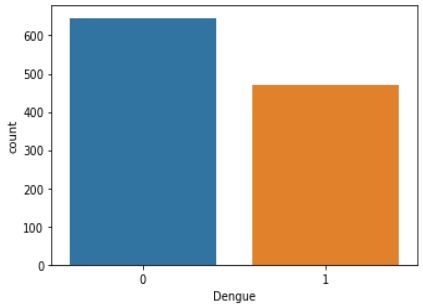

Fig.1. Patient's distribution

3.2 Data Pre-processing

Dataset preprocessing is necessary for making the dataset suitable to use. After preprocess we are able to apply different machine learning method. Real life data are most of the time contain many errors, incomplete and lack of many behaviors. We prepare this data by applying different preprocess method. Computer can only work with numeric value. In this work we convert categorical values to numeric value by using level encoding. We also handle the missing value by using the mean values of that column.

3.3 Principal Component Analysis(PCA)

PCA or Principal Component Analysis which is a dimensionality wane technique that utilized orthogonal variation to reduce a huge set of instances to a tiny set of instances while retaining most of the information present. The procedure works by converting the highly correlated variables of the original dataset to a smaller number of uncorrelated linear instances called principal components. These principal components then account for most of the variance of the original dataset. PCA is very useful when the data has dimensions of 3 or higher since it becomes extremely

hard to make predictions from such large amount of actuality. Furthermore, visualization of high dimensional data is also difficult. PCA solves this problem regarding data visualization as well. Number of principal components is less than or equal to the smaller of the number of original features or the number of observations. Initially, the dataset comprised 1116 observations and 16 useful features, so the maximum number of principal components was 16.

3.4 Train-Test Split

This method is utilized for evaluating the performance of different machine learning algorithms.it is utilized in supervised machine learning algorithms and also utilized in classification and regression problems. In this method the dataset is separated into two parts. The larger part is utilized for training dataset and the smaller part is used for testing the model. In this work we utilized 70% data for trained the model and 30% data for test the model.

3.5 K-Fold Cross Validation

Cross validation is a great way to evaluate performances estimators or machine learning techniques. The dataset is split in to two parts, one for training the estimators and the other to test its performance. After the model is trained, its cross_val_score is calculated, which is the measurement of how good of a learner that model is.K-fold cross validation is a resampling procedure with only one parameter, K, that assesses machine learning algorithms on a finite data. The data sample is to be divided in to K number of groups, hence the name K-Fold cross validation. The idea is to train the algorithm using a little from the data sample, and approximate the performance of the model on unseen data, i.e., data not used in the training phase.

3.6 Algorithms

For implementation the work we use four different machine learning algorithms these are Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN). we applied these algorithms in our dataset for both train-test split and cross validation method. Then compare the result by using different parameter.

Random Forest (RF). Random Forest AI techniques that are utilized for performing both relapse and characterization issues. It is somewhat of an amazing and famous troupe classifier which is utilizing choice tree calculation in a randomize way. It is joined of various choice tree. To assemble each and every choice tree it utilizes sacking method. At the point when we group another item, we got order from each tree as tree vote. What's more, the significant decision in favour of grouping is acknowledged. That is the reason it rather gives more exact result than single choice

tree. What's more, on account of relapse takes the normal of the yields by various trees.

Support Vector Machine (SVM). It is a supervised machine learning techniques utilized in both classification and regression process. An assortment piece capacity is given by SVM to explicit choice capacities, with the chance of indicating custom portions, making them extremely adaptable. Likewise, they are highly retention productive as they utilize subsets of preparing focuses to the choice capacities. For linear classifier, SVM use subset of training data and it is used to represent decision boundary. The hyperplane depends of maximum width of margin so you have to choose most optimal one hyperplane. It is not equally performed well on any unseen example.

Decision Tree (DT). One of the most broadly utilized and effective technique for classification and prediction is the Decision Tree. It is a tree like development anyway changed, for instance its establishment is at the top, included different centre points and branches. At the point when an information comes in the model, decision tree shows the class of that specific info. The meaning of decision node is only test. Test is performed on the quality. To make a decision tree we need a dataset. Then, at that point we need to ascertain the entropy of target worth and indicator quality worth. Then, at that point we will accumulate data of all trait lastly which characteristic has more data it would be root node. In this manner information would be parted and we will get our decision tree and might settle on decision.

K-Nearest Neighbor (KNN). KNN is a kind of Supervised Machine Learning techniques that is normally utilized in classification and prediction problems. Nonetheless, this calculation could likewise be utilized as a predictive regression model. The principle thought of it is to appoint a worth to another information point dependent on the fact that it is so near different information focuses, previously existing in the dataset. By appointing the worth, KNN hence orders that specific information highlight a specific gathering closest to it. As, rather than having a particular preparing stage, this calculation utilizes the information for preparing while at the same time arranging new information focuses, it is named as a Lazy Learner calculation. It is otherwise called a non-Parametric calculation since it doesn't make any presumptions about the fundamental information during classification.

4 Experimental Result Analysis

For finding the expected output I applied percentage split and K-Fold cross validation method. In percentage split method I utilized 70% data for trained the model and 30% data for test the models. On the other hand, in K fold cross validation method where I use the value of K is 10.I also compare the result for

before and after principal component analysis for both percentage split and cross validation method. Where I use 8 principal components. I applied four different machine learning algorithms to this method for finding the dengue disease more accurately. These algorithms are Support Vector Machine(SVM), Decision Tree(DT), Random Forest(RF), K – Nearest Neighbor (KNN). The performance of Machine Learning algorithms is evaluated using several performance metrics. They were accuracy, precision, recall, f1-measure.

4.1 Evaluation Matrix

Accuracy = [(no of true positive + no of true negative) / total number]
Precision(p) = [no of true positive / predicted yes]
Recall(r) = [no of true positive /actual yes]
F1- Measure =[2*(p*r)/(p+r)]

4.2 Compare Results

After obtaining different result in the end of dataset analysis by utilizing machine learning algorithms. We compared the result based on different parameter. In this section we will compare the results and find the best model which is better for prediction of dengue fever disease.

Table 2. Comparison of Evaluation Matrix for Percentage Split

		Precision	**Recall**	**F1-Score**	**Accuracy (%)**
Without PCA	**RF**	0.812	0.760	0.785	82.38
	SVM	0.861	0.830	0.845	87.16
	DT	0.75	0.760	0.755	79.10
	KNN	0.852	0.774	0.811	84.77
With PCA	**RF**	0.804	0.781	0.792	82.68
	SVM	0.863	0.845	0.854	87.76
	DT	0.712	0.697	0.704	74.62
	KNN	0.846	0.774	0.808	84.47

In TABLE 2 we see, that percentage method is used for analyzing the result. Here 70% data is used for training the model and 30% data are utilized for test the model. For Compare the result we applied four different machine learning algorithms which are Support vector machine, Decision Tree, Random Forest.K-Nearest Neighbor. We applied these algorithms for before and after principal component analysis where we use 8 principal components. Here in the table, we see that Support Vector Machine algorithm provides the highest accuracy which is 87.16 % before applying PCA and 87.76% after principal component analysis. Precision, Recall and F1-Score also justify that Support Vector Machine is the best algorithm within all algorithms which we applied in our dataset.

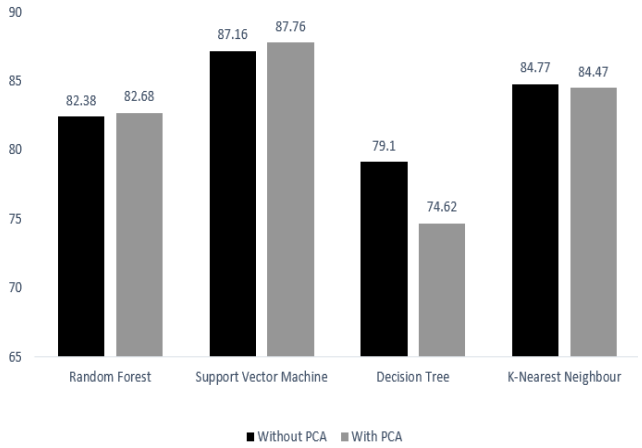

Fig.2. Accuracy result for before and after applying PCA in Percentage Split method

Table 3. Comparison of Evaluation Matrix for 10 fold Cross Validation

		Precision	**Recall**	**F1-Score**	**Accuracy (%)**
Without PCA	**RF**	0.809	0.768	0.788	82.31
	SVM	0.856	0.823	0.839	86.70
	DT	0.715	0.742	0.728	76.31
	KNN	0.823	0.812	0.817	84.73

With PCA	**RT**	0.784	0.768	0.776	80.70
	SVM	0.856	0.823	0.839	86.70
	DT	0.710	0.742	0.726	75.59
	KNN	0.823	0.812	0.817	84.73

In TABLE 3 we see, that k fold cross validation method is used for analyzing the result. Here the value of k is 10. For Compare the result we applied four different machine learning algorithms which are Support vector machine, Decision Tree, Random Forest.K-Nearest Neighbor. We applied these algorithms for before and after principal component analysis where we use 8 principal components. Here in the table, we see that Support Vector Machine algorithm provides the highest accuracy which is 86.70% before applying PCA and also 86.70% after principal component analysis. Precision, Recall and F1-Score also justify that Support Vector Machine is the best algorithm among all algorithms which we applied in our dataset.

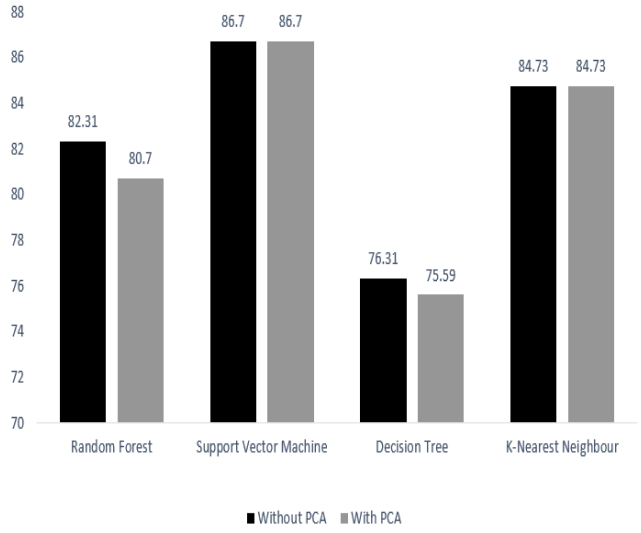

Fig.3. Accuracy result for before and after applying PCA in 10 fold Cross Validation method

5 Conclusion

The target of this research work is to find out a model which can predict dengue fever at an early stage and can reduce the risk of health. If a person provides some symptoms of weakness to the model it can predict that the person has dengue fever or not. For finding that four several machine learning algorithms are utilized which are Support vector machine, Decision Tree, Random Forest, K-Nearest Neighbor. The results are compared by percentage split and cross validation method where 70% data are utilized as trained data and 30% data are utilized for test the model in percentage split method and k-fold cross validation method the value of k is 10.The experimental results show that support vector machine algorithms provide the highest accuracy which is 87.16% in percentage split and 86.70% for 10 fold cross validation before applying PCA.After applying 8 principal component support vector machine also provide the highest accuracy which is 87.76% in percentage split and 86.70% in 10 fold cross validation method. The Precision, Recall and F1-Measure values also performed better in SVM algorithm. In future we can use more related features, Advanced and combined algorithms to improve this model.

References

1. Dengue and severe dengue, https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue, last accessed 2021/09/20.
2. 6,450 hospitalised this year, DGHS worried, https://en.prothomalo.com/bangladesh/6450-hospitalised-this-year-dghs-worried?fbclid=IwAR1oyBDQW7bna-bE37LF4s35wu4oQs970TC_Kg6Qy9jaNe4NDqHcpByVPz0, last accessed 2021/09/20.
3. Dengue wreaking havoc in Bangladesh amid pandemic | Dhaka Tribune, https://www.dhakatribune.com/bangladesh/2021/07/31/dengue-wreaking-havoc-in-bangladesh-amid-pandemic?fbclid=IwAR30RcLO8ntkM1pu5svRxXPsnmujWJpMbEFnRd2lwXGlYNm_aW1CCM-zrdU, last accessed 2021/12/08.
4. Silitonga, P., Bustamam, A., Muradi, H., Mangunwardoyo, W., Dewi, B.E.: Comparison of dengue predictive models developed using artificial neural network and discriminant analysis with small dataset. Appl. Sci. 11, 1–16 (2021). https://doi.org/10.3390/app11030943.
5. Saif Mahmud Khan Dourjoy, A.M.G.R.R., Zerin Nasrin Tumpa, and M.S.: A Comparative Study on Prediction of Dengue Fever Using Machine Learning Algorithm. Presented at the (2020). https://doi.org/10.1007/978-981-15-4218-3_49.
6. Kaur, R.: Demographic Analysis of Dengue Fever Using Data Mining. Int. J. Adv. Res. Comput. Sci. 8, 1233–1237 (2017). https://doi.org/10.26483/ijarcs.v8i7.4565.
7. Shaukat Dar, K., Ulya Azmeen, S.M.: Dengue Fever Prediction: A Data Mining Problem. J. Data Mining Genomics Proteomics. 06, (2015). https://doi.org/10.4172/2153-0602.1000181.
8. Ibrahim, F., Taib, M.N., Abas, W.A.B.W., Guan, C.C., Sulaiman, S.: A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN). Comput. Methods Programs Biomed. 79, 273–281 (2005). https://doi.org/10.1016/j.cmpb.2005.04.002.
9. Mello-Román, J.D., Mello-Román, J.C., Gómez-Guerrero, S., García-Torres, M.: Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. Comput. Math. Methods Med. 2019, (2019). https://doi.org/10.1155/2019/7307803.
10. Laureano-Rosario, A., Duncan, A., Mendez-Lazaro, P., Garcia-Rejon, J., Gomez-Carro, S.,

Farfan-Ale, J., Savic, D., Muller-Karger, F.: Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico. Trop. Med. Infect. Dis. 3, 5 (2018). https://doi.org/10.3390/tropicalmed3010005.

11. Poh, A.H., Mahamd Adikan, F.R., Moghavvemi, M., Syed Omar, S.F., Poh, K., Hisyam Mahyuddin, M.B., Yan, G., Azizah Ariffin, M.A., Harun, S.W.: Precursors to non-invasive clinical dengue screening: Multivariate signature analysis of in-vivo diffuse skin reflectance spectroscopy on febrile patients in Malaysia. PLoS One. 15, 1–22 (2020). https://doi.org/10.1371/journal.pone.0228923.
12. Journal, I.: IRJET- DENGUE FEVER PREDICTION USING CLASSIFICATION TECHNIQUES.
13. Silitonga, P., Dewi, B.E., Bustamam, A., Al-Ash, H.S.: Evaluation of Dengue Model Performances Developed Using Artificial Neural Network and Random Forest Classifiers. Procedia Comput. Sci. 179, 135–143 (2021). https://doi.org/10.1016/j.procs.2020.12.018.
14. Gambhir, S., Malik, S.K., Kumar, Y.: The Diagnosis of Dengue Disease: An Evaluation of Three Machine Learning Approaches. Int. J. Healthc. Inf. Syst. Informatics. 13, 1–19 (2018). https://doi.org/10.4018/IJHISI.2018070101.
15. Fathima, A.S., Manimeglai, D.: Predictive Analysis for the Arbovirus-Dengue using SVM Classification. Int. J. Eng. Technol. 2, 521–527 (2012).
16. Sungheetha, A., Sharma R, R.: Design an Early Detection and Classification for Diabetic Retinopathy by Deep Feature Extraction based Convolution Neural Network. J. Trends Comput. Sci. Smart Technol. 3, 81–94 (2021). https://doi.org/10.36548/jtcsst.2021.2.002.
17. Tripathi, M.: Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM. J. Artif. Intell. Capsul. Networks. 3, 151–168 (2021). https://doi.org/10.36548/jaicn.2021.3.001.