Using machine learning techniques towards predicting the number of dengue deaths in India – A case study

Viswanath Bellie, Madhwaraj Kango Gopal, Govindaraj Venugopal

Professor, Department of Mechanical Engineering, New Horizon College of Engineering, Bengaluru
Professor, Department of Master of Computer Applications, New Horizon College of Engineering, Bengaluru
Assistant Professor, St. Joseph's Institute of Technology, Chennai

Abstract *— Dengue fever has been on the rising end since a few years recently. This has become an alarming sign for the human society today. It is spreading very fast in India and several states have reported multiple admissions and deaths due to this fever. We have analyzed the dengue dataset that was obtained in several states of India and perform a case study to understand the reasons behind the disease getting spread. The objective of this study is to use machine learning techniques in predicting the number of deaths that may arise in the near future. This information will help the government authorities to take necessary steps to decrease the menace caused due to this fever and in turn saving the human population.*

Index Terms *— Machine learning, Prediction,*

I. INTRODUCTION

The rise of dengue fever in India is a serious warning that threatens the very existence of human race if precautionary measures are not taken well in advance. The government and the municipal corporations have been taking a lot of measures and steps but they are not able to stop the spread of this fever. Nowadays, a lot of dengue admissions have been reported in government hospitals and clinics and some percentage of this admission have also resulted in deaths. How are we going to tackle this situation? What methods have to be adopted to be away from this fever? These are some of the questions that need to be answered quickly and if not will erase a lot of human race from this society. Therefore, there is a need to understand this fever in more detail and find out the reasons why this has been spreading so fast. The government has published many datasets regarding this fever and researchers are working on several aspects of this fever. In earlier research works, statistical methods have been used to predict the rise of dengue. In this study, we use machine learning techniques to perform a case study with the dataset obtained from 35 states in India and predict the number of deaths from the admissions done in hospitals.

II. RELATED WORK

Sanjudevi & Savitha [1] have implemented a feature-model construction and done a comparative analysis for improving the predictive accuracy of the dengue disease in several phases. They showed that their classification algorithm SVM gave better predictive accuracy than the decision tree algorithm with the aid of feature selection parameters. Fathima and Manimegalai [2] used the random forest concept by investigating two issues of variable selection for the prediction of dengue. They found a methodology which can be tailor-used for health communication and make better decision support systems for having knowledge of dengue. Iqbal & Islam [3] have proposed a machine learning model to predict the dengue disease. The algorithm that gave the best prediction accuracy was used to build a Dengue prediction tool. Guo et al.[4] have used machine learning algorithms to develop an effective prediction model for dengue. An SVR model showed good results after performing goodness of fit and cross validation tests. Jain et al.[5] used machine learning algorithms to create a dengue prediction model within a specified region which may help in better understanding of the disease and controlling the spread of the disease. Dar & Mehreen [6] used classification techniques like Naïve Bayesian, Random Tree, SMO etc… Data mining techniques were used to compare the performance of all these machine learning algorithms. Kumar et al.[7] investigated through a case study, the various reasons why dengue disease spread in a Tamilnadu, a state of India and gave strong reasons to implement a surveillance system even before dengue spreads in the future. Ceballos-Arroyo et al.[8] developed a machine learning approach for the early detection of dengue disease. They found that an artificial neural network

performed well in the prediction of dengue. Xu et al.[9] developed a forecasting model for dengue based on Long Short Term Memory (LSTM) recurrent neural networks.

III. DATASET USED

The dengue dataset that was used in this study was taken from thirty five (35) states across India. The dataset consists of data
taken from 4 years from the year 2014 to the year 2017. The data consists of both admission of dengue cases in hospitals and deaths that resulted in every year.

IV. RESULTS AND CONCLUSIONS

Several packages of python like matplotlib and seaborn were used for the analysis of the dataset. Several tests were performed on the dataset as follows

A. Histogram of the dataset

A histogram was plotted with the given dataset having the two parameters one the number of admissions and the other the deaths that were reported in a particular year, for the years 2014 to 2017 (till the month of December 2017). The results are represented in the form of a graph in Figure 1.

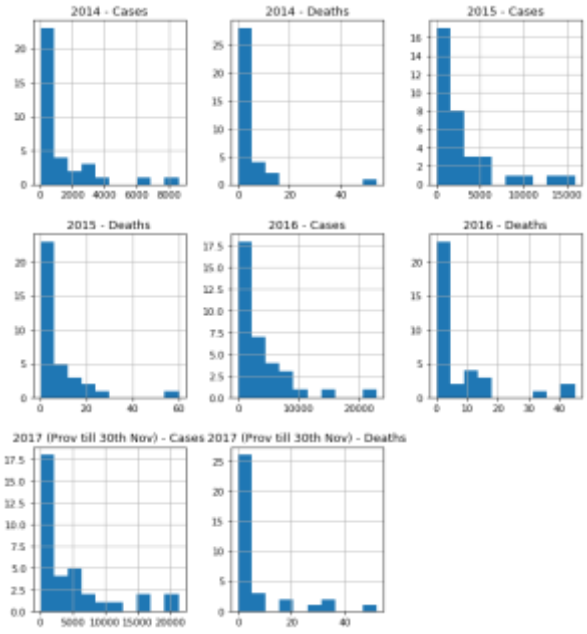

Figure 1. Histogram of the dataset

The histogram was basically constructed to understand if there are any missing values in the dataset that was taken for analysis. From the histogram in Figure 1, it is observed that there are no missing values i.e. all the values are zero.

B. Pair plot of the dataset

The pair plot of the dataset was also plotted. This is shown in Figure 2. To understand the correlation, either positive or negative between the number of dengue admissions and deaths across different states, a pair plot was constructed. This was primarily done to understand the relationships between different variables in the dataset.

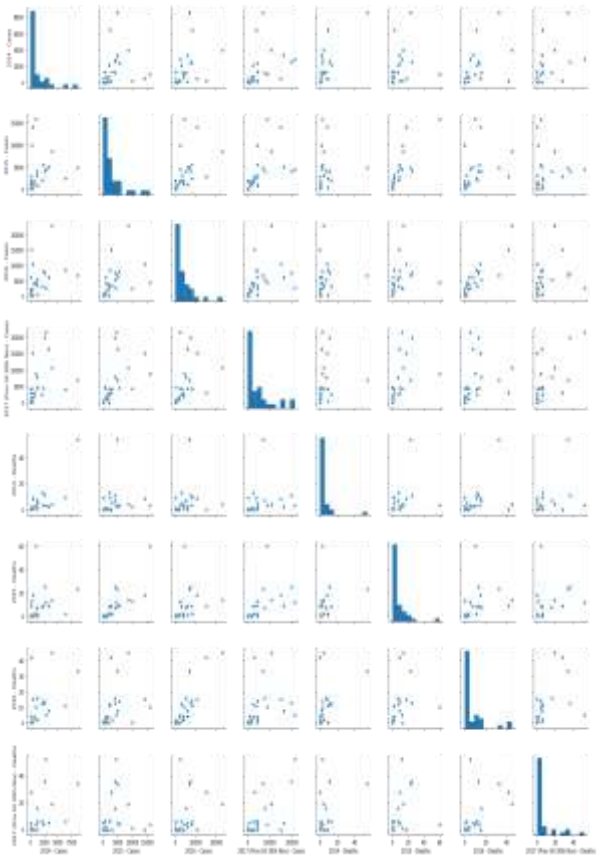

Figure 2. Pairplot of the dataset

C. Detection of Outliers

Outliers are very important artefacts that needs to be taken care while performing different analysis with the given dataset. Outliers are those values that show deviation from the mean of the dataset. When the number of outliers are more in a dataset, the performance of the model will degrade and will not allow us to arrive at a better model. In case the outliers are totally removed from the dataset, the model generated becomes over-fitted, which is not good. As far as possible, the number of outliers in the dataset should be kept minimal. The outliers in both the admission data and the death data for all the four years were analyzed. This was done by using a boxplot structure. The admission outliers are depicted in Figure

3. Subsequently, the death outliers is observed in Figure 4. In both the box plots, there seems be no relationships between admission and deaths. Therefore, it is expected that the model would give a predictive accuracy of say 60-70%.

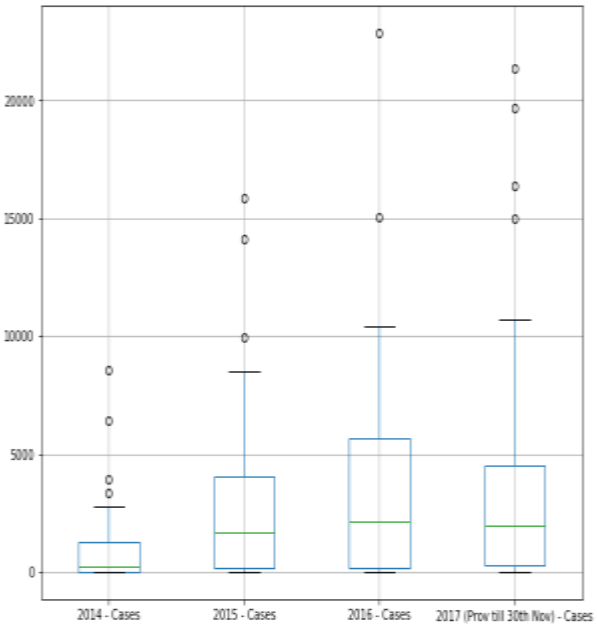

Figure 3. Cases admitted – Outliers detection using boxplot

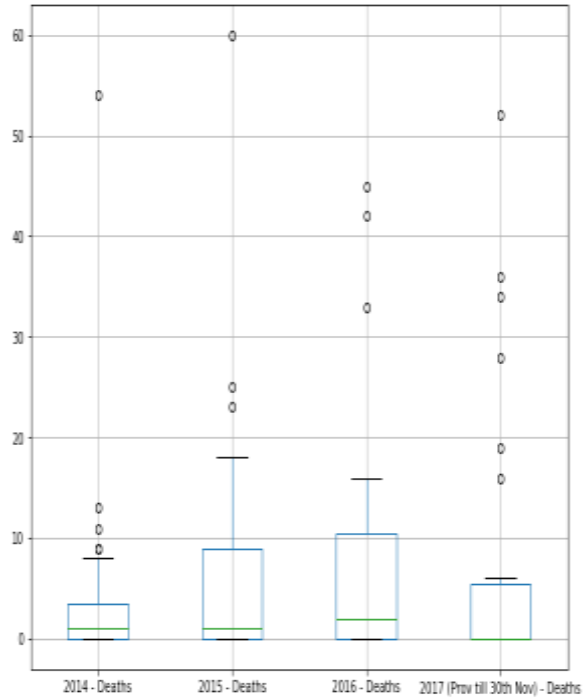

Figure 4. Death Outliers - Detection

D. Distribution of admissions and deaths

The distribution of the different cases that were admitted in the four years was also studied. This is depicted in Figure 5. It is very evident that there is no normal relationship between the two parameters i.e. number of admissions across the four different years. When the dataset forms a normal curve, better insights and conclusions can be arrived at.

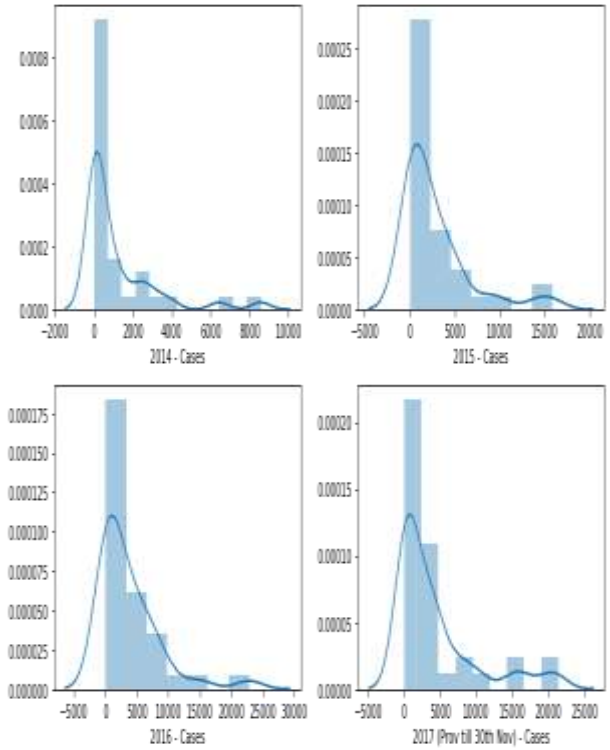

Figure 5. Distribution of Dengue cases admitted from the year 2014 to the year 2017

The distribution of the deaths caused by dengue cases was also studied. This is depicted in Figure 6. As seen in the number of admissions, the number of death cases across the four years are also not showing a normal curve. Therefore, arriving at a best fit model from this dataset would be difficult.

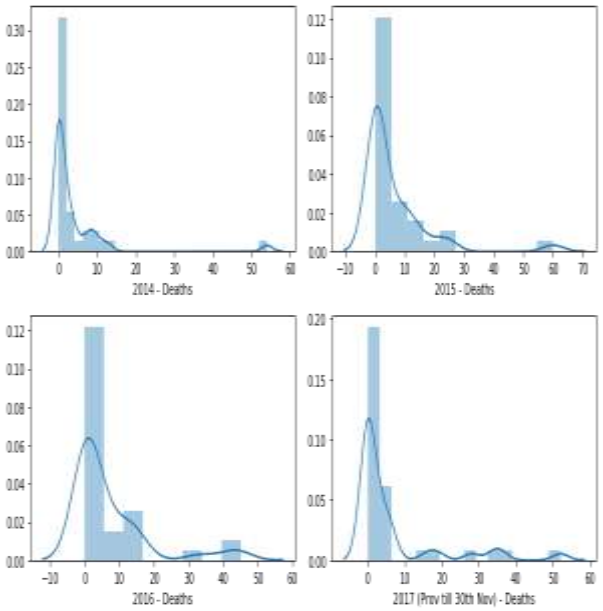

Figure 6. Distribution of Deaths caused through Dengue

E. Correlation Plot

The correlation plot between the number of cases admitted and deaths was also done. The correlation plot is shown in Figure 7. When the correlation plot has the correlated value of more than 0.6, then we say that there is some correlation between the admissions and deaths. From the figure, it is evident that only two correlation values are more than 0.6 and all other values are less than 0.6. Therefore, it is very clear that we will not be able to get a perfect model for dengue disease but will still help us in studying and understanding the relationships to a certain level.

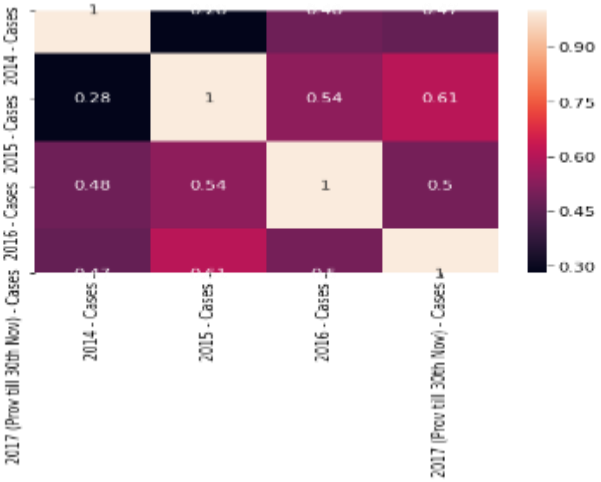

Figure 7. Correlation plot between dengue admissions and dengue deaths

F. Linear Regression

For a better understanding between the relationships between the number of dengue admissions and dengue deaths, linear regression was performed with the data obtained for all the years between 2014 to 2017. The linear regression for the dengue admission data is depicted in the Figures 8 and 9.

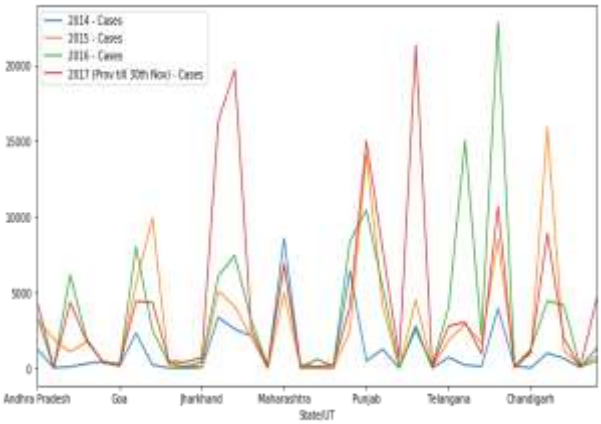

Figure 8. Regression graph for Dengue admissions

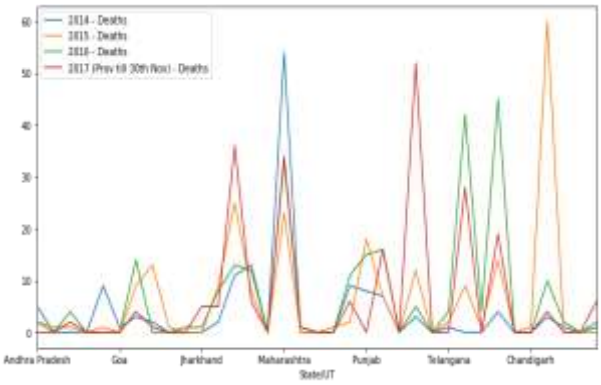

Figure 9. Regression graph for Dengue deaths

G. Autocorrelation Plot

To check for the randomness in the time series data, autocorrelation plots were constructed. The autocorrelation of data values are generated during different time lags. The autocorrelation plot for the dataset is portrayed in Figure 10.

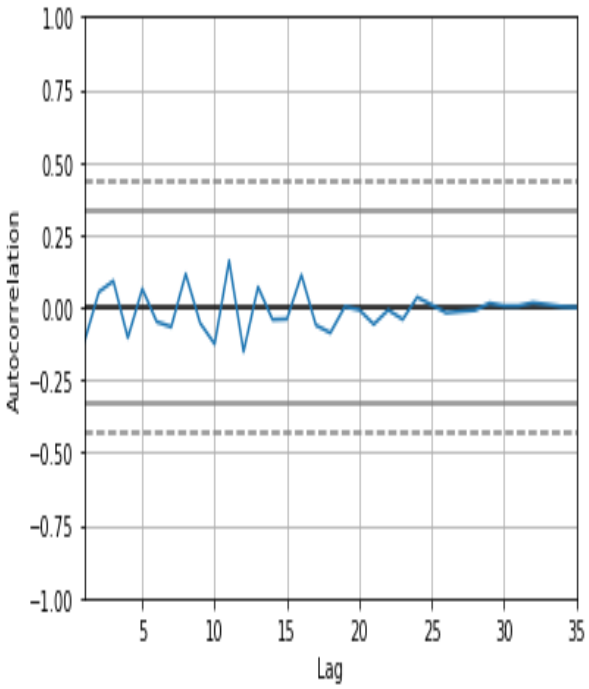

Figure 10. Autocorrelation Plot

H. Scatterplot

To get a better visualization of the data, the admission and death data for one year i.e. the data for the year 2014 was taken and a scatterplot was generated. The scatterplot is shown in Figure 11.

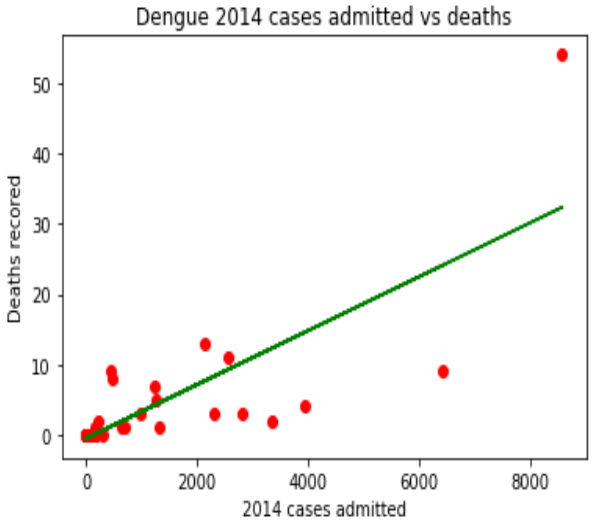

Figure 11. Scatterplot of admissions vs deaths for the year 2014

From the graph, we also found the expected number of deaths if say 'x' patients are admitted into a hospital due to the onset of dengue disease. It was found that if 500 patients get admitted to a hospital, we can expect 1.38 patients to die due to this disease i.e., the average death percentage when compared with the number of admissions in a year. Further, since there was not much relationships or correlation between the admissions and death data across the four years, we did not perform this analysis for the other three years. Even if we had performed these tests for the years 2015, 2016 and 2017, the value would lie somewhere between 1 and 2.

Threats to Validity

In our research, we have only used machine learning algorithms to understand the spread of the disease in different states in India. Much more current existing datasets have to be collected for effective prediction of dengue. Further, death prediction can also be done effectively once more datasets are collected, analyzed with different machine learning algorithms and if required deep learning algorithms can also be used to predict much more effectively in the near future.

V. CONCLUSIONS AND FUTURE WORK

In this case study, we have found that dengue is a dreaded disease that affects the very existence of humanity and if not detected would become a serious threat. We have predicted the death percentage from the dengue dataset collected from several states across India.

As future work, we would be looking at more datasets associated with dengue features and start applying real machine learning and deep learning algorithms to come out with a good model that may help health practitioners to spread the growth of this disease.

We would also like to apply machine learning and deep learning algorithms on the 'Coronavirus' disease and find ways how the spreading of this disease can be stopped.

ACKNOWLEDGMENT

This research was supported by Visveswaraya Technological University, Jnana Sangama, Belagavi – 590018.

REFERENCES

[1] Sanjudevi, R., & Savitha, D. (2018). "*Dengue fever prediction using classification techniques*". International Research Journal of Engineering and Technology, 6(2), 558-563.

[2] Fathima, A. S., & Manimeglai, D. (2015). "*Analysis of significant factors for dengue infection prognosis using the random forest classifier*". Int. J. Adv. Comput. Sci. Appl, 6(2), 240-245.

[3] Iqbal, Naiyar & Islam, Mohammad. (2017). "*Machine learning for dengue outbreak prediction: An outlook*". International Journal of Advanced Research in Computer Science. 8. 93-102.

[4] Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. (2017) "*Developing a dengue forecast model using machine learning: A case study in China*". PLoS Negl Trop Dis 11(10): e0005973. https://doi.org/10.1371/journal.pntd.0005973.

[5] Jain, Raghvendra & Sontisirikit, Sra & Iamsirithaworn, Sopon & Prendinger, Helmut. (2019). "*Prediction of dengue outbreaks based on disease surveillance*", meteorological and socio-economic data. BMC Infectious Diseases. 19. 10.1186/s12879-019-3874-x.

[6] Shaukat Dar, Kamran & Mehreen, Sundas. (2015). "*Dengue Fever Prediction: A Data Mining Problem*". Journal of Data Mining in Genomics & Proteomics. 06. 10.4172/2153-0602.1000181.

[7] Kumar, N. K., Tulasi, R. L., & Vigneswari, D. (2020). "*Investigating dengue outbreak in Tamil Nadu, India*". Indonesian Journal of Electrical Engineering and Computer Science, 18(1), 502-507.

[8] Ceballos-Arroyo, A. M., Maldonado-Perez, D., Mesa-Yepes, H., Perez, L., Ciuoderis, K., Comach, G., ... & Branch-Bedoya, J. W. (2020, January). "*Towards a machine learning-based approach to forecasting Dengue virus outbreaks in Colombian cities: a case-study: Medellin, Antioquia*". In 15th International Symposium on Medical Information Processing and Analysis (Vol. 11330, p. 1133016). International Society for Optics and Photonics.

[9] Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., & Liu, Q. (2020). "*Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method*". International Journal of Environmental Research and Public Health, 17(2), 453.

[10] KUMAR, B. SATISH, and Y. KALYAN CHAKRAVARTHY. "*PREDICTION OF OPTIMAL TORQUES FROM GAIT ANALYSIS APPLYING THE MACHINE LEARNING CONCEPTS*." International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) 9.4 (2019):685-698

[11] Chandra, Dimple, and Partibha Yadav. "*Prediction Of Software Maintenance Effort On The Basis Of Univariate Approach With Support Vector Machine*." International Journal of Computer Science And Engineering (IJCSE) 3.3 (2014): 83-90.

[12] Sathawane, N. K. S., and Pravin Kshirsagar. "*Prediction and Analysis of ECG Signal Behaviour using Soft Computing*." IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET) 2.5 (2014): 199-206.

[13] Dhevarajan, S., et al. "*SPR_ SODE Model for Dengue Fever*." International Journal of Applied Mathematics & Statistical Sciences (IJAMSS)", 2.3 (2013): 41-46.

[14] PATEL, AJAY M., A. PATEL, and HIRAL R. PATEL. "*COMPARATIVE ANALYSIS FOR MACHINE LEARNING TECHNIQUES APPLIANCE ON ANOMALY BASED INTRUSION DETECTION SYSTEM FOR WLAN*." International Journal of Computer Networking, Wireless and Mobile Communications (IJCNWMC) 3.4 (2013): 77-86.

[15] SRIVASTAVA, SRISHTI, et al. "*ANALYSIS AND COMPARISON OF LOAN SANCTION PREDICTION MODEL USING PYTHON*." International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) 8.2 (2018):1-8

[16] Durgabai, R. P. L., and P. Bhargavi. "*Pest Management using Machine Learning Algorithms: A Review*." International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) 8.1 (2018): 13-22.

[17] Gawaly, A. M., M. Ghazy, and S. Abdou. "*Early Prediction for Common Complications of Liver Cell Failure Using Fecal Calprotectin Concentration.*" International Journal of Medicine and Pharmaceutical Sciences (IJMPS) 6.1 (2016): 131-138.