JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF HEPATITIS A DISEASE USING MACHINE LEARNING TECHNIQUES

[1]Nagireddi Surya Kala, [2]G L Narasamba Vanguri, [3]Juvvala Sailaja
[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor
[1]Information Technology,
[1]Aditya College of Engineering and Technology, Surampalem, India

Abstract: Stagnated water is the root cause of the outbreak of Water-borne and Vector-borne diseases. During floods excess water is stagnant over time. Due to this stagnant water, food and water is contaminated. Individuals who consume the contaminated water can contract Hepatitis A. Predicting the occurrence of Hepatitis A disease involves analyzing various risk factors and medical data to identify individuals or populations that are more likely to contract the disease . Communicable diseases need to be diagnosed in the early stages to save the lives of victims. This paper focuses on prediction of Hepatitis A, a communicable disease using Machine Learning and Deep Learning Models. Random Forest, Multilayer Perceptron and Probabilistic Neural networks models are used to analyze the hepatitis data set for disease prediction. Multilayer Perceptron model shows better performance when compared to other models.

Keywords **- Flood water, Waterborne, Vector-borne, Multilayer Perceptron, Probabilistic Neural Networks, Random Forest**

I. INTRODUCTION

Flooding increases the likelihood of exposures to Water-borne and Vector-borne diseases. Water sources are contaminated with floodwater. Waterborne diseases are infections transmitted through water. Consumption of drinking water contaminated with bacteria and viruses leads to waterborne diseases in human beings. These diseases also propagate with an intake of food washed with floodwater and with open wounds or direct eye contact. Intake of food washed with contaminated water causes infections. Vector-Borne Diseases are caused by vectors which are acting as an agent for transmitting the virus from human to human and animals to human. Stagnant water during floods increases the expansion of vector habitats. Waterborne and Vector-Borne diseases and their virus are shown in Fig1.

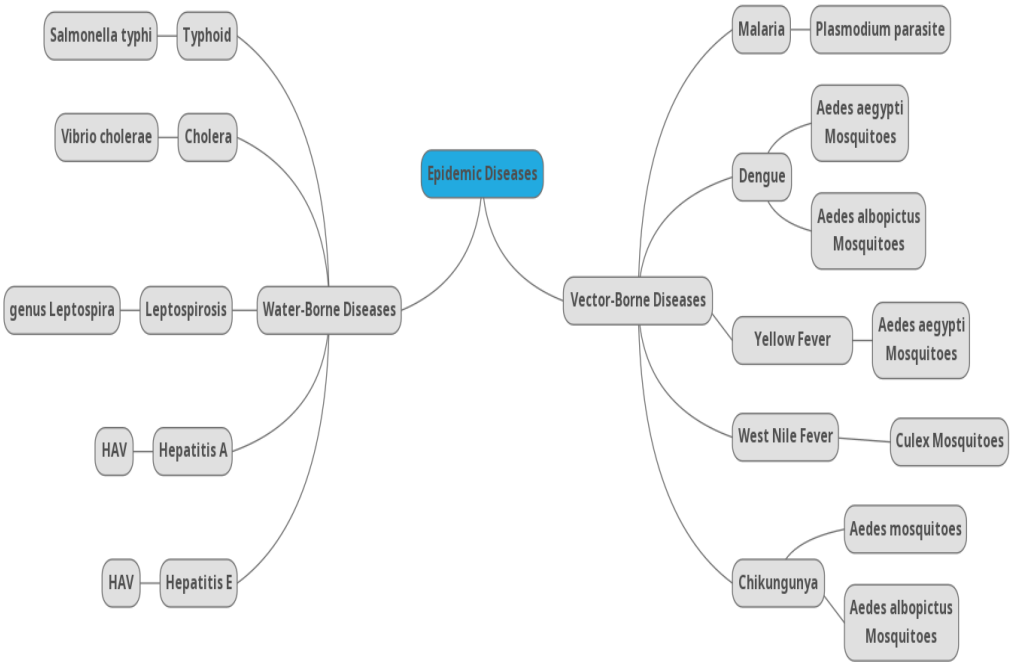

Fig 1 Epidemic Diseases

This paper focuses on hepatitis A waterborne disease infected due to flood-related sewage contamination of water sources. When the water source is infected with feces of an infected person, then water is contaminated with HAV Virus. Consumption of water with the HAV virus leads to Hepatitis A disease. .An Early Diagnosis system is essential to diagnose these diseases accurately in

the early stages. The proposed Method focus on prediction of Hepatitis A disease when flood related sewage is contaminated in various water sources which is being implemented using multilayer Perceptron , Random Forest, and Probabilistic Neural Networks models. Section 2 discuss the literature survey focusing on prediction of Communicable diseases using Machine learning and Deep Learning techniques. Section 3 explains the proposed methodology for prediction of hepatitis-A disease. Section 4 shows results and discussions. Section 5 is followed by conclusion.

II. Review of Literature

Rashi Bhardwaj conducted a relative study of different classification Algorithms for Prediction of Liver diseases. Among these Techniques XGBoost showed better accuracy in classifying patients according to Liver Disorder[1]. Letícia M. Raposo performed comparative analysis for HIV Drug resistance prediction. Compared with linear regression and classification tree Random Forest is best for analysing genomic data.[2] Earth Observation data is collected to get environmental attributes since the Vector population data is necessary to prevent transmission of virus. Machine Learning models provide better results when compared to Linear statistic models for modelling the temporal distribution for adult female mosquitoes [3]. A Baldominos conducted a systematic review on how Computational intelligence is used for the prediction of infectious disease. Early Diagnosis of these diseases helps the patients to recover soon and stop spreading of the diseases to others [4]. Kalansuriya discussed about the factors affecting Dengue disease such as environment and population.Machine Learning Techniques are compared to analyse the spreading of Dengue[5,6] .

D Aabhas proposed an alerting system for prediction of propagation of vector-borne diseases based on weather and other influencing factors using Machine Learning Algorithms[7]. Yang Chen performed Machine Learning Classification for prediction of early stage of liver fibrosis. Diagnosis of liver fibrosis at an early stage minimizes the risk of developing liver cancer[8]. Chiara Garattini discussed about controlling the spreading of infectious diseases in emergency outbreaks using Big data Analytics.[9] Bigger data provided to Deep cascade Forest model shown better performance for prediction of water quality[10]. S. K. Lakshmanaprabu proposed an IOT based health care system. The e-health data extracted from patients infected with different diseases is classified with Random Forest Classifier [11]. Machine Learning Models helps in recognizing sex and gravity of flying mosquito species from the backscatter signals recorded using an optical sensor for monitoring of Mosquito population[12]. K Nearest Neighbour, Naive Bayes, and Extreme Gradient Boost Models are applied on historical dataset in order to establish a relationship between malaria outbreak and weather conditions[13].

Sangwon proposed a method for prediction of infectious diseases using deep learning models with big data and social media data. ARIMA model has shown better performance when compared with other models for predicting infectious diseases[14]. Big data ,AI and machine learning models are used for prediction , management and improvement of Vector borne diseases. [15]. Neural Networks are used to identify the association between flood parameters and water quality using RadarSat-2 [16]. Use of Artificial intelligence methods for analysis of infectious diseases and surveillance data to the concerned Authorities reduces the impact. Proposed a polynomial regression model of sixth degree as the best model for forecasting COVID 19 for next six days in India.[17] Prediction of epidemic diseases proactively using Machine Learning, Deep Learning and improved mathematic modelling the growth of epidemic disease is predicted proactively[18]. Kamel Boulos discussed the role of Artificial Intelligence and Geographic Information System in Health care[19].

III. RESEARCH METHODOLOGY

3.1 Data Set:

Hepatitis dataset is taken as historical data for analysing the severity of hepatitis disease based on the attributes age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver_big, liver_firm, spleen_palable, spiders, ascites, varices, bilirubin,alk_phosphate, sgot, albumin, protime, histology. This is a binary classification problem where Target variable labelled severity=dead or live. Pre-processing techniques such as Data normalization and discretization are used to eliminate missing and null values for accurate predictions. Hepatitis dataset used for risk analysis of Hepatitis A disease is downloaded from the Kaggle website.

3.2 Prediction of hepatitis A disease using Multilayer Perceptron:

Multilayer Perceptron is a supervised fully connected feed forward neural network which consists of multiple layers. As shown in Fig 2, a multilayer perceptron model is developed layer by layer with an input layer, required hidden layers, and one output layer. Hepatitis dataset consists of 142 records. Out of these records, for training the model 75% (106 records) are used and for testing the model 25% (36 records) are used. Input attributes are fed to the input layer along with initial weights, followed by hidden layers and output layer. Weights are adjusted to reduce the error between the actual and predicted values of the target variable for better predictions. First Layer is the Input dense layer with activation function as ReLU. The second layer is one hidden Layer with activation function as ReLU. More than one hidden Layers can be added to improve the model. The output layer, here only one node is added with sigmoid as an activation function for binary classification. This is a binary classification model that is used to classify whether the severity of the patient is live or die. This model is compiled with loss function as binary_crossentropy , optimizer as adam, and metrics as accuracy. Then the model is fitted with a training set and the number of epochs is 150 and batch size as 10. The trained model is used for prediction using the test case.

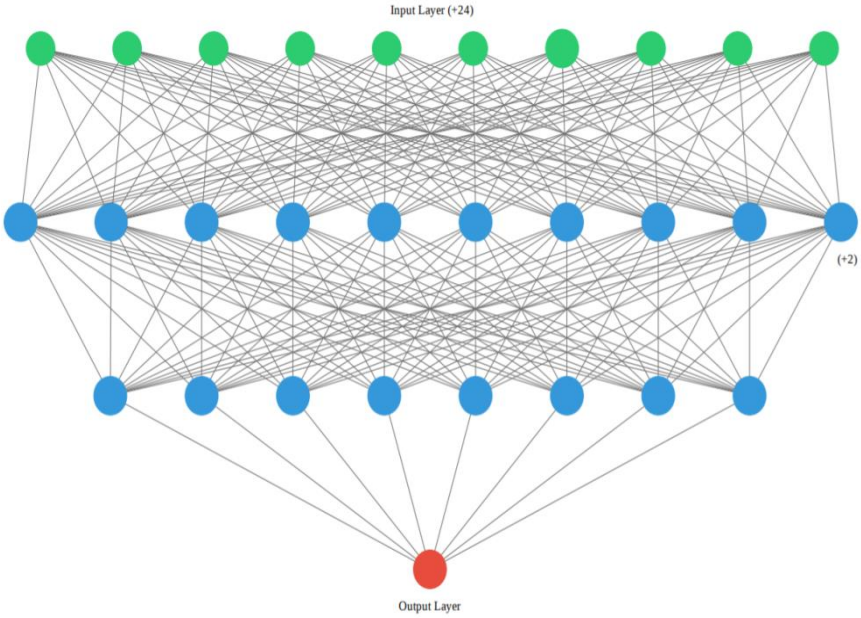

Fig 2 Multilayer Perceptron model

3.3 Prediction of hepatitis A disease using Random Forest:

Random Forest Classifier model is trained by constructing a large number of decision trees by choosing subsets of training dataset as shown in Fig 3. This classifier works as an ensemble approach by combining the results of all the decision trees based on the majority voting scheme.

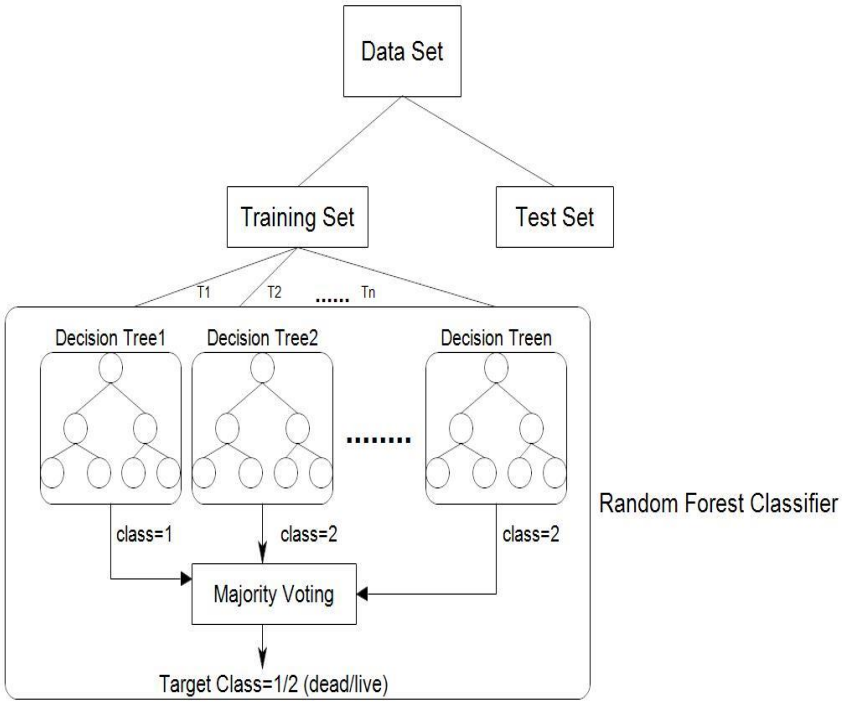

Fig 3 Random Forest Classifier

Input Dataset is divided into training and test sets. Sampling procedure is applied to generate samples using subsets of attributes from the training data set. Decision trees are constructed from the subsets generated. Training records are classified based on the aggregation of votes from all the decision trees using Majority voting scheme. After the model is trained new test cases are fed to all the decision trees and the output class severity is dead or live is based on a majority vote.

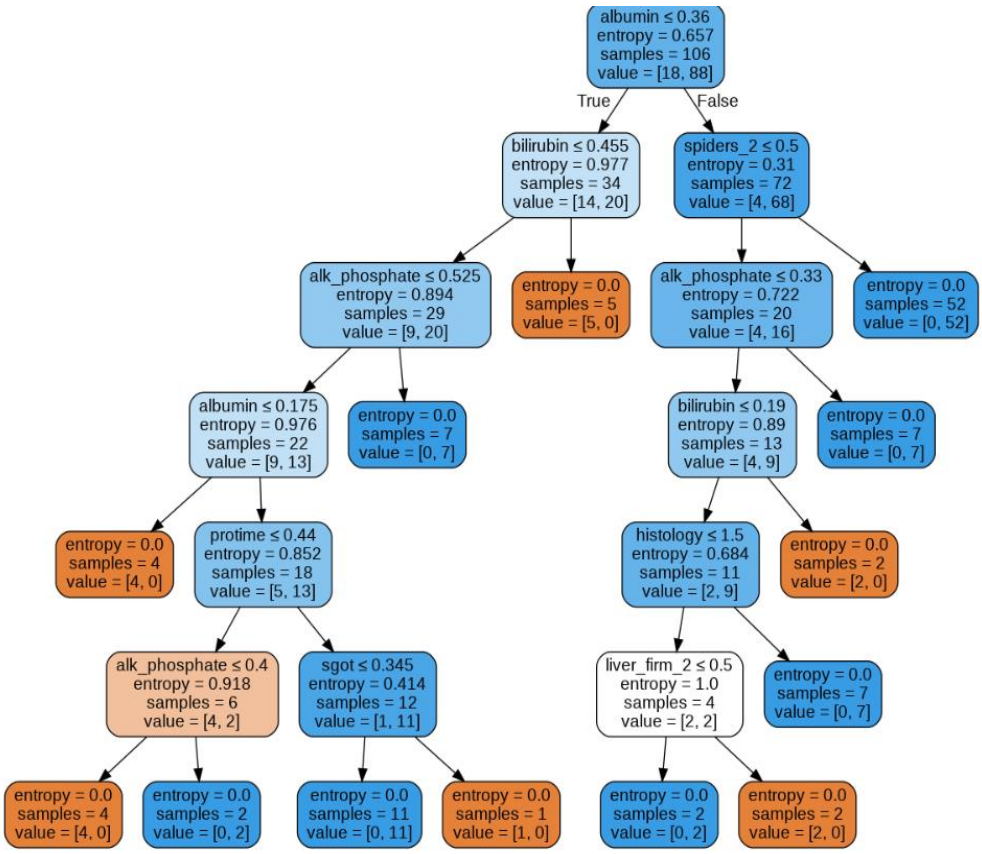

Fig 4 Visualization of Decision Tree for disease prediction

Fig 4 shows the patterns found in the decision tree. All the rows begin with a single bin at the root of3 the tree. A subset of features is considered to construct a different decision trees. The features selected show how data rows are divided into the most useful way with entropy measure. The root node contains the most informative condition albumin<=0.6. If this is true, the left branch of the tree is traversed to get 34 samples of value=[14,20] i.e there are34 samples of class= 'die'. The remaining 72 samples of a total of 106 samples go to the right sub tree. The splitting continues until the sub tree ends with a bin with only one class or entropy measure reaches to 0.0

3.4 Prediction of hepatitis A disease using Probabilistic based neural networks:

A PNN is a classification algorithm based on kernel discriminant analysis in which the operations are organized into a multi-layered feed forward network is used for predicting the severity of the hepatitis disease as a live or dead target attribute. In this method, a probability distribution function is used for each class for estimating the class probability of test data based on Bayes rule based on the highest posterior probability.

Fig 9 shows the Probabilistic Neural Network model for hepatitis A disease prediction. The Attributes x1,x2… xn vectors in the training dataset are fed to the Input Layer. The Pattern layer consists of radial basis neurons for each training sample. The Euclidian distances from the input vector to the training input vectors forming a pattern neuron for each category. The summation layer sums the inputs provided by the pattern layer from the corresponding category. The output Layer is a decision layer that selects the class having a higher posterior probability.

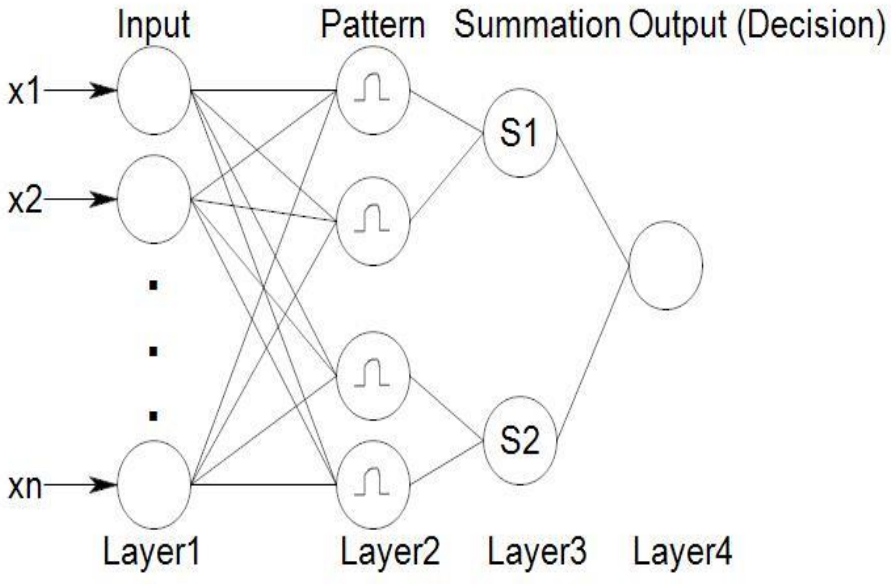

Fig 5. Probabilistic Neural Network model for disease prediction

IV. RESULTS AND DISCUSSION

4.1 Comparison of Model Performance using Accuracy Metrics : The Performance of multilayer Perceptron, Random Forest, and Probabilistic Neural Network models are evaluated using Accuracy, Area under the curve, Recall and Precision, and tabulated as shown in Table1.

Table No.1: Performance Metrics

Algorithm	Accuracy	Area under the curve	Recall	Precision
Multilayer Perceptron model	94.444444	85.9375	0.968750	0.96875
Random Forest	94.444444	75.0000	0.941176	1.00000
Probabilistic Neural Network	86.111111	81.2500	0.965517	0.87500

4.2 Performance of Classification Models using the Receiver Operating Characteristic curve: The receiver operating characteristic curve is plotted as shown in Fig 5 .This Graph shows the performance of all the classification models at all classification thresholds. The graph is plotted with X-Axis as false positive and Y-axis as True Positive.The area under the curve depicts the two-dimensional area measurements under the ROC curve, showing the aggregate performance measure at all the thresholds.

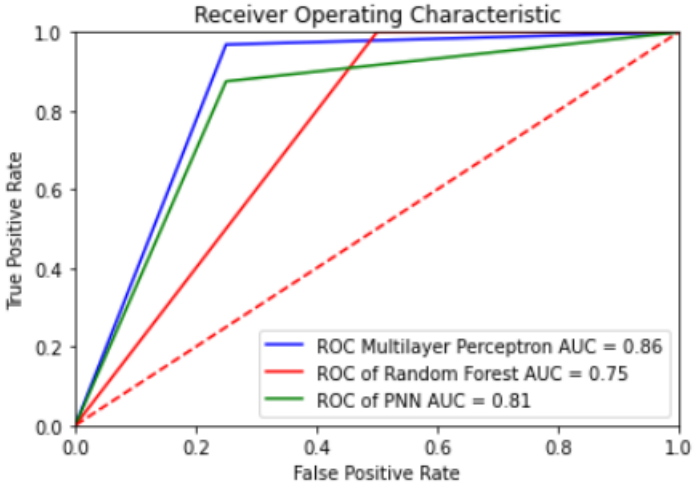

Fig 5 Receiver Operating Characteristic curve

IV. Conclusion

Hepatitis A is a contagious viral disease that infects humans from mild to severity. Hepatitis symptoms are taken as input attributes in the dataset with target attribute class severity. The target variable with class=1 indicates that the person will not survive and class=2 indicates the person will survive. At first Data Analysis and pre-processing, steps are performed on the input dataset for accurate predictions. Multilayer Perceptron, Random Forest, and Probability Neural Network models are trained on the Hepatitis dataset to predict the severity of Hepatitis disease. Results show that the Multilayer Perceptron is proved to be the best with an accuracy of 94.4% and the area under curve as 86%.

REFERENCES

[1]. R. Bhardwaj et al. " A Comparative Study of Classification Algorithms for Predicting Liver Disorder" Intelligent Computing Techniques for Smart Energy Systems, Springer,2019

[2]. L.M.Rapso et al. "Random Forest Algorithm for Prediction of HIV Drug Resistance" Science, Technology, Engineering, Agriculture, Mathematics & Health, Springer, 2020

[3]. O.Mudele et al. "Modeling the Temporal Population Distribution of Ae.aegypti Mosquito using Big Earth Observation Data" IEEE Access 2020

[4]. A.Baldominos et al. "Predicting Infections using computational Intellignece" IEEE access 2020.

[5]. Carvajal et al. "Machine Learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitian" BMC Infect Dis 2018

[6]. C.S Kalansuriya et al. "Machine Learning-Based Approaches for Location Based Dengue Prediction" Advances in Intelligent Systems and Computing. Springer, 2020

[7]. D Aabhas and Prabhishek Singh "Comparative Analysis of Epidemics Alert System using Machine Learning for Dengue and Chikungunya" International Conference on Cloud Computing, Data Science & Engineering 2020

[8]. YangChen et al. "Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B", Computers in Biology and Medicine 2017

[9]. Garattini et al. "Big Data Analytics", Infectious Diseases and Associated Ethical Impacts, 2019

[10]. Chen Kangyang et al. "Comparative Analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data" Water Research, 2020

[11]. Lakshmana et al. "Random Forest for big data classification in the Internet of things using optimal features", Int. J. Mach. Learn. & Cyber, 2019

[12]. A P.Genond et al. "A comparision of supervised machine learning algorithms for mosquito identification from backscattered optical signals" Ecological Informatics 2020

[13]. K Godson et al. "Predicitng Malarial outbreak using Machine Learning and Deep learning Approach" International Conference on Information Technology 2018

[14]. C Sangwon et al. "Predicting Infectious Disease using Deep Learning and Big Data" International Journal of Environment Research and public Health 2018

[15]. Debra P.C et al. "Big data-model integration and AI for vector-borne disease prediction " An esa open access journal 2020

[16]. P Yomwan "A study of waterborne diseases during flooding using Randarsat-2 imagery and a back propogation neural network algorithm" Geomatics, Natural Hazards and Risk 2015

[17]. Ramjeev Singh et al. "Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India" International Journal of Information Technology, 2020

[18]. T Shreshth et al. "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing" Internet of Things 2020

[19]. Kamel Boulos et al. An overview of GeoAI applications in health and healthcare Int J Health Geogr, 2019