© June 2021| IJIRT | Volume 8 Issue 1 | ISSN: 2349-6002

Efficient Covid 19 Forecasting for worldwide countries using ANN

Yogini Jawale[1], Akshay Thakare[2], Aaditya Shinde[3], Govind Waghmare[4], Mrs.A.G.Said[5]
[1,2,3,4,5] *Computer dept, AISSMS IOIT, Pune*

***Abstract* - The sudden emergence of the Covid-19 Pandemic has been one of the most problematic scenarios experienced by the global community in the recent decades. This has been especially devastating due to the large death toll and increasing economical strain due to the successive lockdowns and restrictions in place to combat the epidemic. This has been highly problematic to contain the spread of the pandemic which his highly unpredictable. The losses that have been incurred by the governments across the world have been due to problems arising by the lack of effective prediction of the Covid-19 infection rates. The predictions would allow the health sector to be better prepared for the infection number which can provide a significant boost to their efforts. Therefore, an effective approach for the prediction of Covid-19 infection rates has been illustrated in this research article. The presented approach utilizes Linear Regression along with Artificial Neural Networks and Fuzzy Classification. The extensive experimentation has been performed to determine the performance of the approach which has led to satisfactory results.**

***Index Terms* - Linear Regression, Artificial Neural Network, and Fuzzy Classification.**

I.INTRODUCTION

COVID-19, a new coronavirus, caused an epidemic in Wuhan, China, in December 2019. It quickly expanded to over 200 nations throughout the world following its inception. The number of verified COVID-19 infected patients is expanding at an exponential rate throughout the world. COVID-19 has infected over 17 million individuals to far, with over 0.7 million people dying as a result. As a result of the lack of particular immunizations to prevent the disease from spreading further, numerous nations have entirely shut down their everyday operations. Lockdown has been implemented across the country. It controls the spread of the disease to some extent; yet, it has had a significant impact on the national and global health economy. The COVID-19 pandemic has mostly impacted small, medium, and big businesses, who are experiencing issues such as decreased demand, no exports, a scarcity of raw materials, and transportation and supply chain interruptions.

Thousands of COVID19 patients come every day in need of rapid help, which is frequently unavailable. As a result, it is critical to develop a technological tool that can estimate the number of infected individuals and foresee the worst-case situations, such as the economy collapsing and the healthcare system collapsing. The impact, on the other hand, is determined by the characteristics of the afflicted people, such as social connections, personal economic and educational levels, and government resources to deal with the crisis. Because there is no vaccine for COVID-19 and the disease is communicable, the number of afflicted persons is rising at a quicker rate. The tests that measure the Coronavirus to determine the presence of illness, are to be examined more since the range of symptoms of positive cases has been rising since the virus's discovery on the planet. As COVID-19 has achieved pandemic status and the number of patients continues to rise at an exponential pace, widespread diagnostic testing is critical in determining and controlling the spread of this rapidly spreading illness.

Fuzzy Rule-Based Systems (FRBS) are a type of artificial intelligence that uses fuzzy notions to make decisions. The goal of the methods is to describe knowledge as a set of fuzzy rules. For tackling difficult real-world issues, FRBS has been proposed [11]. The FRBS's performance is determined by its membership functions and rule base. Li- Xin Wang and Jerry M. Mendel proposed the Wang and Mendel (WM) fuzzy rule learning approach in 1992 [12].

Most of the preceding models take into account time-based data, which is generally monotonic and has predictable behavior that can be explained by a quiet basic model. A distinct problem linked to COVID-19 will be addressed in this research. The goal of this study is to create a model that predicts the number of confirmed cases caused by COVID-19 in the selected country based on data from several countries. These data are not time-based in theory, and their behavior is typically random. Linear Regression and ANN supervised-based approaches are utilized and evaluated in terms of their prediction abilities to develop the best forecasting model.

In this research article related works are mentioned in the section 2. The proposed technique is deeply narrated in the section 3. The experimental evaluation is performed in section 4 and whereas section 5 concludes this research article with the scope for future enhancement.

II.LITERATURE REVIEW

M. Qjidaa et al. proposed the Transfer Learning principle and pre-trained architectures such as DenseNet121, VGG16, VGG19, InceptionResNetV2, Xception, MobileNet, InceptionV3, Deep Learning algorithms were used to categorize three classes COVID-19, Pneumonia, and normal. The proposed system can identify distinguishing characteristics in chest X-ray pictures and used the seven models to create an ensemble model that outperformed all others [1]. In general, the suggested global model is made up of two modes. The input photos are transformed into descriptor vectors using a pre-trained start mode. Another option consists of many classifiers that are tightly linked together, with each classifier producing its own pre-processed output. The forecast with the highest score is the one that the global system will keep when it exits. The suggested model, in more detail, is made up of three primary processes. With a test accuracy of 99 percent, an f1-score of 98 percent, a precision of 98.60 percent, and a sensitivity of 98.30 percent, the final classifier outperformed the competition.

P. Gawade et al. states that the healthcare facilities are few in underdeveloped nations, and the ecosystem is ill-equipped to fight the epidemic in industrialized nations, the pandemic might kill thousands of people. The best method to deal with the scenario is to take proactive measures to comprehend an individual's health based on his medical state, history, and other factors [2]. The presented methodology shows that the machine learning models can learn quickly using the knowledge supplied by the embedded media dataset, corroborating the assertions of subject matter experts in the field of personification. The pertained models can deliver the safety score with greater than 95% accuracy. It is clear from studies that one may obtain a safety measure based on the medical data. These danger variables will serve as a warning, allowing them to prepare depending on preconditions and attain personification.

A. A. Prakash et al. used data from Indian states as well as data from other hard-hit nations across the world in the presented methodology. To begin, they classified the Indian states into three groups based on how badly they've been affected. Top-5 most affected, top-5 moderately impacted (middle), and top-5 least impacted states are the three categories [3]. This classification is based on the author's bold premise that states in the same category have comparable dissemination patterns and, as a result, require similar preventative actions from the government and healthcare providers. Aside from that, the forecasts tend to be smoother when similar-hit states are combined. Furthermore, for the sake of simplicity, they have omitted elements such as the intensity of the lockdown enforced, the availability of healthcare facilities, and have based system modeling only on the infection statistic.

N. S. Wibowo et al. presented a covid19 prediction system using logistic regression and random forest combined with the ensemble algorithm, namely bagging. The first step in applying machine learning is to provide data, which is separated into two parts: a training dataset and a testing dataset. The training dataset is where a machine learning model is built. The machine learning model is tested using the testing dataset. After the data is separated, each preprocessing dataset goes through a procedure that tries to remove unwanted noise from the data in machine learning preprocessing, which is done at this point in the form of text. Deleted words, punctuation connections, and unneeded words are all part of text preparation. The feature extraction method generates documents that may be used to train machine learning algorithms [4]. TF-IDF is a commonly used algorithm extraction feature. This approach works by calculating the relative frequency of a set of events. by comparing the inverse of the word on a certain document. The word ratio in the dataset compared to the entire dataset or Corpus. Model fits are used to construct a machine learning model. The model was then used to anticipate testing a unique dataset from which the outcomes could be assessed. If the evaluation is not satisfactory, the fits model is repeated until the assessment is satisfactory. The logistic, decision tree, and random forest algorithms were used to test regression. After the results were good, two algorithms would be the best to perform ensemble use bagging classifier, and model machine learning has become the model that can be employed. Confusion matrix and accuracy value are used to evaluate machine learning models.

M. F. Jojoa Acosta et. al., presented a machine-learning-based strategy for predicting likely confirmed Covid19 cases in six different American nations, which might be useful for planning during the pandemic's containment stage. For this goal, they offer two conventional regression models and compare their performance. Multilayer Perceptron and Support Vector Machine are the proposed models [5]. It's worth noting that they chose these machine learning techniques since the amount of accessible data is limited, with an average of 78 registers per nation corresponding to confirmed cases gathered day by day during the pandemic's progress from the start date of measurement to May 25. Improvements in performance indicators for the MLP model of Chile, Mexico, and the United States were achieved using the suggested optimization strategies. Pearson's correlation coefficient indicates that the data trend is continuing and MAE and MPE both perform admirably. On the contrary, In the same measures, SVM excels in Brazil and Colombia as well as Peru.

E. F. Ohata et al. presented a technique for categorizing an X-ray as being from a healthy patient or a COVID-19-affected patient. First, they go through the datasets and then explained the feature extraction, which is based on transfer learning. Their findings demonstrate that extracting features with CNNs, applying the transfer learning approach, and then categorizing these features with consolidated data is a good way to go. Using machine learning approaches to categorize Xrays is a good idea [6]. Thus, it does not replace a medical diagnosis since a more thorough investigation could be done with a larger dataset. Under those circumstances, their work contributes to the possibility of an accurate, automatic, fast, and inexpensive method for assisting in the diagnosis of COVID-19 through chest X-ray images.

D. Haritha et. al. presented a technique for predicting COVID-19 infected patients from patient chest X-ray pictures using Googlenet is given. They showed how transfer learning can be used to predict COVID-19 for the first time. This model may be used to diagnose COVID-19 from patients' chest X-rays quickly and accurately. Their trained model has a testing accuracy of 98.5 percent and a training accuracy of 99 percent. Primary health care providers in rural areas where experienced practitioners are unavailable might utilize this computerized COVID-19 prediction method [7]. This study might be expanded to create a large collection of X-rays of COVID and non-COVID patients with various Pneumonia illnesses to enhance specificity and sensitivity. The technology may be combined with the Internet of Things (IoT) to help medical practitioners even more, as the disease is spreading at a quicker pace and no vaccine has been identified yet.

A. Jarndal et. al. suggested GPR-based model was benchmarked against an ANN model to forecast the number of fatalities caused by the new COVID-19. The influence of age, the number of smokers, and diabetic individuals is increasing. The number of people who have died as a result of this extremely contagious sickness is increasing [8]. The ANN-based model could not be capable of accurately anticipating behavior due to the extreme type of randomness present in the data. Provided adequate and balanced data, ANN efficiency can dramatically improve. The efficiency of ANN can be substantially improved. However, as compared to ANN, GPR has been demonstrated to be significantly more effective. The GPR system can be easily simulated and projected because of its probabilistic and non-parametric character. In modeling/forecasting COVID-19 data, GPR is shown to be more successful than ANN.

V. K. Gupta et. al., based on a data-driven approach, provides machine learning methods. Using the existing data, their technique predicts the number of persons infected with COVID-19 in the following days [9]. This research provides a methodology that can accurately predict the number of new COVID-19 instances, allowing management to plan for their handling. Feature selection is used throughout the model building process to choose the most relevant features out of all the characteristics. It decreases the prediction model's complexity. The authors used the R programming language to do feature selection using the random forest significance algorithm. The aforementioned procedure, whose input parameters include all the features of the COVID-19 cases dataset in India, is used to calculate the classification model features. Confirmed, death and recovered cases are the qualitative metrics. Machine learning approaches do not include any new information from other models or template structures in this case. The accuracy of each model is assessed. The consistency of the random forest model, which delivered an almost linear performance in the prediction of all these scenarios, is measured using K-fold cross-validation.

P. Cihan[10] presented the WM technique which is used to anticipate the number of confirmed Covid-19 cases. The study's data set spans the periods of March 27, 2020, and July 28, 2020, as announced by the Turkish Ministry of Health. The number of instances was estimated using the following input variables: total number of critical care patients, the total number of intubated patients, number of daily tests, number of daily recovered patients, and number of daily death. R-square, mean absolute error (MAE), and root mean square error (RMSE) statistical criteria were employed to assess the FRBS method's prediction ability.

III PROPOSED METHODOLOGY

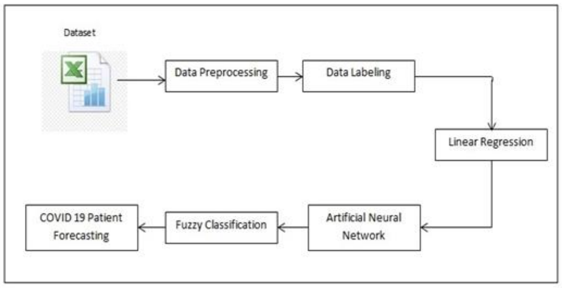


Figure 1: Proposed model System Overview

The presented technique for the purpose of achieving the effective and useful Covid-19 infection rate prediction has been depicted in the figure 1 given above. The presented technique has been elaborated in a stepwise manner in the section given below.

Step 1: Dataset collection, preprocessing and Labeling – The system for the prediction of the Covid-19 infection rates have been achieved through the use of an extensive dataset extracted from the URL: https://www.kaggle.com/imdevskp/corona-virus-report

The extracted dataset is provided as an input to the proposed methodology for performing the predictions. The dataset is interfaced in the form of a workbook that is read by the system through the use of the JXL API. This Library provides the java code the ability to interface with the workbook file.

The dataset contains a collection of attributes out of which only certain attributes are necessary for our implementation that can be useful for achieving the accurate prediction. The extracted attributes are the country/region name, confirmed cases and the 1 week change in the cases which are stored in the form of a double dimension list. The country/region is effectively converted in to integer format through the labeling procedure. Through the labeling process an integer is allotted to a particular unique region for easier interfacing and appended to the list which is provided as to the subsequent step.

Step 2: Linear Regression – In this step of the procedure the previous step provides the input in the form of a double dimension list. This list contains the selected attributes that are preprocessed and labeled. The regression on these attributes is achieved through the use of the linear regression approach.

The linear regression achieves a regression between an independent and a dependent variable that changes constantly. These values are referred to as the x and a y value, where x is the independent value and y is the dependent value. This is given by the equation 1 given below.

$$Y=Mx + B \qquad (1)$$

This equation is used to achieve the value regression by the value of m which is the slope and b which is the intercept is unknown. These values are achieved by providing the dependent and independent values, which are the 1 week change attribute values and the confirmed number of cases respectively. These values are provided from the input preprocessed and labeled dataset in the form of a list for intercept and slope measurement through the equations 2 and 3 given below.

$$M = \frac{\mathrm{N}\sum(xy)-\sum x\sum y}{\mathrm{N}\sum(x^2)-(\sum x)^2} \qquad (2)$$

$$B = \frac{\sum y - M\sum x}{\mathrm{N}} \qquad (3)$$

Where:

x = Independent variable (1 week change in infections)
y = Dependent variable (Confirmed infections)
M = Slope or Gradient (how steep the line is)

B = the Y Intercept (where the line crosses the Y axis)
N= Size of the array
Y=Intercept value

The values once achieved can be utilized in the equation 1 to achieve the value of the dependent variable. This is done by taking into account the user input from the system for a particular region. The values of the x or the independent variable are provided and the respective values of y are achieved as a regression. These values are utilized further for the process of Covid-19 infection rate estimation.

Step 3: Artificial Neural Network – The preprocessed and labeled list containing the required attributes from the dataset is provided as an input to this module for the evaluation of the hidden and output layers. The Artificial Neural Networks execute in the form of layers which are, namely, input layers, hidden layers and output layers. For the detection of the Covid-19 infection rates the input values of the two attributes, the confirmed infections and the one week change in infection for a particular region is provided as an input. These values are utilized along with the assigned random weights for the purpose of creating the two hidden layers. The two weights for each of the attributes is used and combined with the bias weights. The two hidden layers are evaluated through the implementation of the RELU activation function. The achieved hidden values are provided to the output layer for the evaluation of the output error probability. The two output layer utilizes the two hidden layer values along with the weights and the bias weights to achieve the output layer values. These values are utilized in the equation 4 given below, along with the two target values, T1 as 0.01 and T2 as 0.99 to achieve the error probability rate.

$$Error\ Probability = \sum \frac{1}{2} (\mathrm{T}_0 - O_L)^2 \text{-----} \ (4)$$

Where,
T = Target Values
OL = Output Layer Values

This error probability rate is then appended at the end of the row in the double dimension list of the respective region. This is repeated for all the regions. The resultant list is provided for sorting in the ascending order of the error probability rate. The error probability rate is inversely proportional to the accuracy of the probability achieved which will be used for further classification.

Step 3: Fuzzy Classification – The error probability list obtained in the previous step is utilized as an input in this step of the procedure. This list is already sorted in the ascending order, whose length is extracted and divided by 5. This value is used to segregate the list into 5 clusters, with the varying fuzzy crisp values of VERY LOW, LOW, MED, HIGH and VERY HIGH going from top to bottom. These clusters are also ranked from 1-5 respectively.

The user input is also taken in consideration in this step of the procedure, and the corresponding region is searched in the clusters for its position according to the respective label.

Once the position of the input region is attained, it is divided by the total number of clusters which is 5, the obtained value is then combine with the regression value of the respective region achieved in the previous steps. This resultant value is then added to the one week change value of the respective region, if the rank of the region is more than or equal to 3, otherwise the value is subtracted from the one week change to achieve the accurate covid-19 infection rate estimation. The process of prediction label extraction from the fuzzy crisp set can be seen in the below mentioned algorithm 1.

ALGORITHM 1: Prediction Label Through Fuzzy Crisp values

```
//Input : Fuzzy Crisp Set FCSET , CL  Input Country Label
//Output: Prediction Label PDL
Start
    PDL =0
    for i=0 to Size of FCSET
       SG = FCSET [i] [SG = Single Cluster]
           for j=0 to Size of SG
             row= SG [j]
           RCL= row[0] [ RCL = Row Country Label]
              if( RCL =CL), then
                PDL= i+1
                break
              end if
         end for
    end for
return PDL
Stop
```

IV RESULT AND DISCUSIONS

The proposed methodology for the prediction of Covid-19 infection rate prediction has been developed through the use of Java programming language. The NetBeans IDE has been utilized to facilitate development of the proposed methodology. The development machine is equipped with 4GB of RAM and 500GB of Hard Disk which is supplemented by an Intel Core i5 processor. The JXL API is being used for the purpose of enabling the interfacing of the dataset in a workbook format.

Experimental evaluation has been performed to allow an in-depth assessment of the prediction procedure for the presence of any errors. The error assessment derives the preciseness of the prediction that can be highly useful in describing the accuracy of the Covid-19 prediction system. The RMSE or the Root Mean Square Error performance metric has been used to extract the error of the prescribed prediction model.

The RMSE technique utilizes two continuous and correlated entities that are correlated to determine the error between these two variables. The variables utilized in our implementation are the expected Covid-19 infection rate predictions and the obtained Covid-19 infection rate predictions. The equation utilized to achieve the error are illustrated in the equation 5 given below.

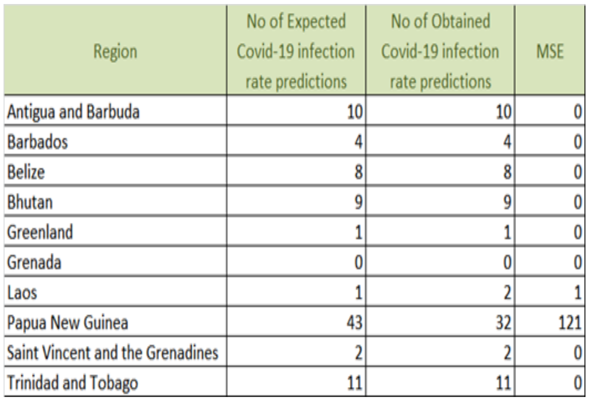
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1,i} - x_{2,i}\right)^2}{n}} \qquad (5)$$

Where,

$\sum$ - Summation

$(x_1 - x_2)^2$ - Differences Squared for the summation in between the expected Covid-19 infection rate predictions and the obtained Covid-19 infection rate predictions

n - Number of samples or Trails



Region	No of Expected Covid-19 infection rate predictions	No of Obtained Covid-19 infection rate predictions	MSE
Antigua and Barbuda	10	10	0
Barbados	4	4	0
Belize	8	8	0
Bhutan	9	9	0
Greenland	1	1	0
Grenada	0	0	0
Laos	1	2	1
Papua New Guinea	43	32	121
Saint Vincent and the Grenadines	2	2	0
Trinidad and Tobago	11	11	0

Table 1: Mean Square Error measurement

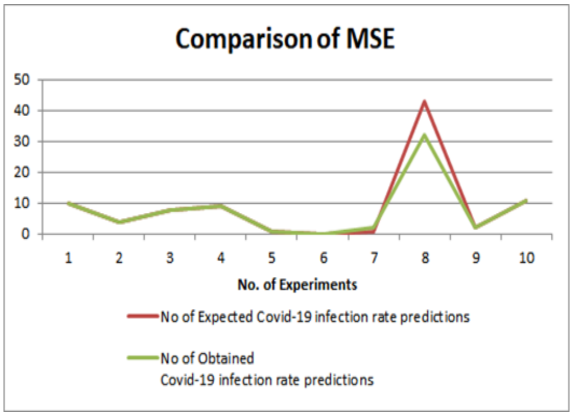


Figure 2: Comparison of MSE in between No of expected Covid-19 infection rate predictions V/s No of obtained Covid-19 infection rate predictions

The experimental evaluation and its outcomes have been depicted in the table 1 given above. The values obtained in the table are used to draw a line graph given in the figure 2 given below. Through close examination of the graphical representation and the tabulated values, we can come to a conclusion that the error achieve in this process of prediction is minimal. A collection of 10 experiments have been conducted with varying regions have been performed to achieve the assessment of MSE or Mean Square Error.

The assessment results declare that the error achieved in the prediction system is acceptable and reasonable. The error for the prediction is usually present in the prediction models which are work on a real world data. There are a number of different scenarios that affect the Covid-19 infection rate predictions. The achieved MSE and RMSE values of 12.20 and 3.49 respectively are highly satisfactory and describe an accurate deployment of the Covid-19 infection rate prediction model.

V. CONCLUSION AND FUTURE SCOPE

The presented technique for the evaluation of the Covid-19 infection rates has been elaborated in this research article. The approach utilizes the Covid-19 dataset as an input which is then preprocessed and the relevant attributes for the prediction purposes have been selected. The region attribute is also converted in to an integer format through the use of the labeling approach. The preprocessed and labeled dataset is provided to the next module for regression evaluation through the use of Linear Regression. These values are then stored and will be used later in the process. The user input for the region is taken and the Artificial Neural Network is initiated for Hidden layer and Output layer Estimations. This is realized for achieving the error probability of the prediction. These values of regression and error probability of the Covid-19 infection rate prediction are provided to the Fuzzy Classification module for the purpose of classification. The classification approach determines the infection rates after one week of the current infection numbers. These values are then provided to the user through the graphical user interface. The presented approach for the prediction has been subjected to experimentation and achieved MSE and RMSE values of 12.20 and 3.49 respectively, which have been highly satisfactory.

The future research direction can be in the direction of achieving the Covid-19 prediction approach in the form of a web application of ease of use by the end users. This approach can be inducted on district wise dataset for more accuracy.

REFERENCES

[1] M. Qjidaa et al., "Early detection of COVID19 by deep learning transfer Model for populations in isolated rural areas," 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), 2020, pp. 1-5, DOI: 10.1109/ISCV49265.2020.9204099.

[2] P. Gawade and P. S. Joshi, "Personification and Safety during the pandemic of COVID19 using Machine Learning," 2020 4th International Conference on Electronics, Communication, and Aerospace Technology (ICECA), 2020, pp. 1582-1587, DOI: 10.1109/ICECA49313.2020.9297555.

[3] A. Prakash, P. Sharma, I. K. Sinha, and U. P. Singh, "Spread & Peak Prediction of Covid-19 using ANN and Regression (Workshop Paper)," 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), 2020, pp. 356-365, DOI: 10.1109/BigMM50055.2020.00062.

[4] N. S. Wibowo, R. Mahardika, and K. Kusrini, "Twitter Data Analysis Using Machine Learning to Evaluate Community Compliance in Preventing the Spread of Covid-19," 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020, pp. 1-4, DOI: 10.1109/ICORIS50180.2020.9320816.

[5] M. F. Jojoa Acosta and B. Garcia-Zapirain, "Machine Learning Algorithms for Forecasting COVID 19 Confirmed Cases in America," 2020 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2020, pp. 1-6, DOI: 10.1109/ISSPIT51521.2020.9408742.

[6] E. F. Ohata et al., "Automatic detection of COVID-19 infection using chest X-ray images through transfer learning," in IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 1, pp. 239-248, January 2021, DOI: 10.1109/JAS.2020.1003393.

[7] D. Haritha, N. Swaroop and M. Mounika, "Prediction of COVID-19 Cases Using CNN with X-rays," 2020 5th International Conference on Computing, Communication, and Security (ICCCS), 2020, pp. 1-6, DOI: 10.1109/ICCCS49678.2020.9276753.

[8] A. Jarndal, S. Husain, O. Zaatar, T. A. Gumaei, and A. Hamadeh, "GPR and ANN-based Prediction Models for COVID-19 Death Cases," 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), 2020, pp. 1-5, DOI: 10.1109/CCCI49893.2020.9256564.

[9] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," in Big Data Mining and Analytics, vol. 4, no. 2, pp. 116-123, June 2021, DOI: 10.26599/BDMA.2020.9020016.

[10] P. Cihan, "Fuzzy Rule-Based System for Predicting Daily Case in COVID-19 Outbreak," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2020, pp. 1-4, DOI: 10.1109/ISMSIT50672.2020.9254714.

[11] L. A. Zadeh, "Information and control," Fuzzy sets, vol. 8, pp. 338-353, 1965.

[12] L.-X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," IEEE Transactions on systems, man, and cybernetics, vol. 22, pp. 1414-1427, 1992.