Journal of Advanced Medical and Dental Sciences Research

@Society of Scientific Research and Studies ***NLM ID:*** *101716117*

Journal home page: www.jamdsr.com doi: 10.21276/jamdsr Indian Citation Index (ICI) Index Copernicus value = 100

(e) ISSN Online: 2321-9599; (p) ISSN Print: 2348-6805

# Original Research

# A Hybrid model for COVID-19 Prediction based on Regression Techniques

[1]Deepti Malhotra, [2]Gurinder Kaur Sodhi

[1]Research Scholar, [2]Associate Professor, Desh Bhagat University, Punjab, India

***ABSTRACT:***
As of Jan 2023, 67 crore COVID cases have been identified causing 68 lakh death worldwide. It has been more than three years since its first onset but till present date, we still live in fear of getting infected by it. It has been declared as one of the deadliest virus in the history of the mankind. In addition to different precautions and vaccines, researchersare now focussing on robust prediction and forecasting models for an accurate prediction. The present work proposes a Hybrid model based on Regression techniques. A time series-based data set from Kaggle; consisting of death count, confirmed and recovered cases has been used as an input for this. The prediction has been carried in three steps mainly: Data collection, Feature Extraction and Regression. The presented model has lower values of MSE, MAE and RMSE in comparison to other models, with values as 15675791, 3218.17 and 4045.45.

**Keywords:** Prophet, Random Forest, XGBoost regression, Voting regression

Received: 02 August, 2023 Accepted: 06 September, 2023

**Corresponding Author:** Deepti Malhotra, Research Scholar, Desh Bhagat University, Punjab, India

**This article may be cited as**: Malhotra D, Sodhi GK. A Hybrid model for COVID-19 Prediction based on Regression Techniques. J Adv Med Dent Scie Res 2023;11(10):1-6.

## INTRODUCTION

In December 2019, cases of severe respiratory sickness were reported across different parts of the world. The SARS-CoV2 coronavirus, also known as COVID-19 was the cause of this sickness. Midway through January, the number of cases of this disease surged quickly as the virus had spread outside China. The virus, subsequently spread in the entire world. After through and comprehensive researches, a detailed analysis of the virus has been done. Its mode of transmission and preventions both have been understood, vaccines too have been developed. However, the virus too has been mutating with time. In addition to all the precautions being adopted, need of predictive models is sort of mandatory now. Hence many statistical and neural network based predictive and forecasting models have been proposed and implemented in recent years [1]. However, for successful and accurate prediction, transparent reporting and multiple model evaluations are required. It is possible to analyse the state and trend of Coronavirus using a variety of time series prediction methods. The main goal of the forecasting is to create a pattern that can accurately represent the relationship by analysing past views of an irregular variable and forecasting its future values. This idea can be applied in various situations [2]. Over the years, there have been several attempts and studies in building and enhancing time series models. Since past few years many researchers have been using time series techniques with statistical and Machine Learning (ML) influences to predict the number of active Covid-cases. Leptospirosis modelling and its relationship with precipitation and temperature are common examples for these. Some of the most commonly used time series-based methods are: Facebook Prophet, ARIMA, Random Forest, XGBoost, TBAT, HWAAS, N-Beats and Deep AR [3]. The time series model known as the ARIMA is one that is frequently used for economical applications. The statistical characteristics of this model lead to its adoption and fame. It uses Box-Jenkins' approach during the training phase and can use a variety of available exponential based smoothing methods. Such models remove all the high-frequency noise from the time-series data before extracting the local patterns. These, are majorly described by three separate input parameters: a, d and q, where:

$a$ is the lag order or the amount of lag observations included in the model $d$ is the number of times differentiation of the raw observations is done, $q$ is the size of the moving average window required for calculating the mean.

For instance, let a = 0, d = 2 and q = 2, then the ARIMA forecast value for $\widehat{Y}_t$ is equal to:

$\widehat{Y}_t = 2Y_{t-1} - Y_{t-2} - \theta_1 e_{t-1} - \theta_2 e_{t-2}$ (1)

Here, $\theta_1$ and $\theta_2$ are moving average coefficients, $e_t$ is observation's predicting error, at time t. This prediction technique uses significant autocorrelation between data into account [4]. Moreover, ARIMA may be used with variety of dataset types and still yield the best results. Another time series based popular model that can also deal with trends and weather variations is the Holt-Winter Additive model (HWAAS). In this, the in-trend forecast includes a seasonal component to predict knowledge with an occasional or regular component in a better manner. This can be used in place of ARIMA, however experiential studies have shown that the method is less accurate while calculating averages in comparison to former models. Although, HWAAS model is effective when used on weather- based data with course and seasonal characteristics that remain consistent over time and hence can provide accurate estimate while representing fluctuations [5]. Some of the practical obstacles encountered during the deployment of this technique are: selecting initial attributes, its susceptibility to unusual events or anomalies, choice of smoothing metrics and the standardisation of climatic index.

The next method is trigonometric seasonal combination of Box-Cox transformation, ARMA error and trend component (TBAT). This method is based on the use of trigonometry-based functions in simulating non-integer cyclical frequencies and Box-Cox translations for normalising dependent variables and enabling various types of non-linearities. This method is effective for addressing a range of time-series-related problems and is mathematically effective in high probability assessing of complex weather-based problems. This division is carried to identify and separate the seasonal components which are hidden in the time series plot.

Another time series approach called Prophet has been created by Facebook to address concerns with business time series [6]. This method involves the creation of a model that is easily decomposable and includes trend, seasonality and holiday components as different trend elements. This model based on regression techniques can be easily adjusted and it performs well while using its default values. It also helps the algorithm in choosing the components for forecasting and applying the required adjustments, subsequently. For implementation, mainly two systems are used here: a saturating growth model and a piecewise linear model. A pattern that requires nonlinear growth at a saturation point of carrying capacity is employed to carry out the growth forecasting. The most common application of this is population-based patterns paradigms. This is a piece-wise model of stable development value, which creates a useful and consistent arrangement which helps in predicting the saturation points.

Another flexible model called Prophet, is used to capture seasonality of periodic effects that depends on the Fourier series [7]. Deep AR is a probabilistic forecasting method based on the principal of Autoregressive Recurrent Neural Networks (ARNN). For accurate predictions, appropriate probabilities are combined with nonlinear data transformations which are trained using a Deep Neural Network (DNN). The model is built on the foundation of deep learning (DL) for time series data, using a RNN structure and largely addresses the problem of probabilistic prediction. Using an in-depth heap of completely linked layers and forward and backward surviving connections, the times series approach N-Beats uses deep neural architecture [8]. In general, models of such type are implemented using traditional techniques like seasonality and trend level approach. These stacks are made up of various blocks connected by the leftover connections. This dual residual stacking provides the trend component in conjunction with forecast theory and is subsequently removed from the input window in prior to being sent into the seasonality stack. As a result, seasonality and partial trend forecasts are decoupled, and this separate conclusion is used to add an alluring layer of relevance to the pattern.

## LITERATURE REVIEW

Tingzhen Liu, et.al (2020) devised an algorithmic strategy for predicting autocorrelation sequences [9]. The model creates a temporal series of location-based reference functions that undergoes a full pandemic condition cycle. The reference function ensembles are transformed into various observable forms, along with other fields of imperfect data. The modified functions here; accurately forecast the series values that were concealed in other areas. The work demonstrates the use of data from additional autocorrelation prediction-based exogenous variables to enhance the model's output.

Sina Ardabili, et.al. (2020) carried out a review of use of an Artificial Neural Network along with the Gray Wolf Optimizer approach for forecasting the COVID-19 pandemic. A global dataset has been used for this for undertaking verification, training, and testing procedures [10]. This paper used Mean Absolute Percentage Errors (MAPE) and R values for evaluation of the results. The proposed method generates a MAPE of 6.23-13.15 for training and 11.4%-11.5% for testing and validation phases respectively.

Raghavendra Kumar, et.al. (2021) proposed a time series models based on John Hopkins Coronavirus Resource Centre data [11]. A combination of three time series models have been used —ARIMA, MA and AR to predict the active cases of Corona virus in India. The new model may serve as a useful tool for creating and implementing decision-making methods in various scenarios. The predictions made by this model were quite accurate and they may potentially be improved and confirmed using more ML and DL algorithms.

Andi Sulasikin, et.al. (2020) made the prediction of active cases of Covid-19 for the Indonesian population by using a combination of two time series models: ARIMA and Holt's exponential smoothing [12]. Results showed that for estimating the potential number of active Corona Virus Disease cases in 2019 in the Indonesian capital, the ARIMA model had highest R-squared ($R^2$) and lowest MSE and RMSE. The employed system delivered amazingly good performance and produced accurate predictions that backed data-driven policy in the healthcare sector.

Zehua Yu, et.al (2021) constructed novice framework, called IT-GCN to review the cases of Covid-19 outbreak [13]. The model is a combination of ARIMA and GCN for data modelling based on nodes in a graph that reflected the severity of the outbreak for numerous regions. In order to find the relationship in the data, this work builds the graph nodes with the vectors using ARIMA parameterization. The results demonstrated that the proposed model performed better in predicting short-term daily cases of corona illness.

Radu Beche, et.al (2020) developed an auto-encoder forecasting model for predicting the COVID positive cases across many European countries [14]. The results of this model were utilised by officials for skilfully dividing resources and in planning preventive actions to lessen the impact of the infection in their respective area. The strategy that was provided could be used to prevent the COVID-19 sickness from spreading. Notably, the suggested model did not take into account the context of any other types of data, such as mitigation strategies, travel effects, etc. and instead just employed mathematical data.

Leonardo Sestrem de Oliveira, et.al (2021) demonstrates the use of ANN model to forecast the COVID positive cases and fatalities over the course of following week using time series data for three nations [15]. The iterations demonstrates that ANN model accurately anticipated the total number of deaths and positive cases. In general, a comprehensive test set for an ANN resulted in a 50% higher MSE than a random test set for the same network.

## RESEARCH METHODOLOGY

Time series analysis tool is the foundation of this research. Some of the parameters that are used to evaluate the model's performance are Mean Squar Error, Mean Absolute Error and Root Mean Square Error [16-18].

1. Mean Squared Error: - The value indicates the closeness of the regression line to group of points. Let the predicted values on the test instances be $f_1$, $f_2$,$f_3$.....$f_n$ and the actual values are $y_1$, $y_2$,....$y_n$ ,then the MSE is:

$$\mathrm{MSE} = \frac{(f_1 - y_1)^2 + \cdots + (f_n - y_n)^2}{n} \qquad (2)$$

2. Mean Absolute Error: - This is a measure of arithmetic averages of the absolute errors. It is normally used in the measure of forecast error.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|(3)$$

Here, $f_i$ is the prediction and yi is the true value.

3. Root Mean Square Error: - The squared root of the mean squared error yields the root mean square error (RMSE). This parameter gives more weight to the larger and infrequent errors.

**Following are the steps carried for the time series data analysis:**

1. Dataset Collection and Pre-Processing: - The dataset for the present work has been acquired from Kaggle. This dataset comprises of global cases in form of time series. Three such categories have been included in the target set in the form of: recovered cases, confirmed cases and death cases.
2. Feature Extraction and Regression: - For prediction analysis, multiple models like Prophet,XGBoost and Random Forest have been used. The breakdown of the time series model into trend, seasonal and holiday components is displayed as:

$$\mathcal{Y}(t) = g(t) + s(t) + h(t) + \mathcal{E}(t)(4)$$

Here, t is the passage of time, $\mathcal{Y}(t)$ is a time series and g(t) is the trend term which is a function for fiting the time series' aperiodic portion. Seasonal change is represented by the function s(t), holiday effect is denoted by h(t), which depicts changes brought on by exceptional events like holidays and $\mathcal{E}(t)$ is the noise part of the time series i.e. unpredictable random variations**.** A type of Gradient Boosting Decision Tree strategy, that may be applied to both classification and regression problem is the XGBoost method. By including more weak learners, the Gradient Boosting seeks to rectify the residuals of each weak learner. Here, accuracy is greater for a single student when numerous learners are combined to make the final prediction. The objective function of XGBoost is defined as the sum of loss function and a regular term [19].

$$\mathrm{Obj}\Theta = L(\Theta) + \Omega(\Theta)$$
(5)

Each region's output value is $c_i$ and is expressed as:

$$\min_{j,s}\left[\min_{c_1}\mathrm{Loss}(y_i, c_1) + \min_{c_2}\mathrm{Loss}(y_i, c_2)\right] (6)$$

If the space is partitioned into 'f' elements i.e.$R_1, R_2, \ldots, R_f$, then the regression tree's equation is:

$$f(x) = \sum_{f=1}^{F} c_f I(x\epsilon R_f)(7)$$

Several regression trees are created in the end. If the set of regression trees is $\{T_1, T_2, \ldots, T_s\}$ and when input is fresh data, every tree will have a different prediction value. In the end, the final prediction is expressed as the average value of the prediction results of all the trees, given by [20,21]:

$$f(x) = \frac{1}{s}\sum_{i=1}^{s} T_i(x)(8)$$

Overall, the regression model of Random Forest reduces the variance of the model by integrating each tree's forecast. The majority voting determines the outcome. The plurality voting method is fairly straightforward and allows each voter to choose just one option. The winner is determined by which candidate receives the most votes. The likelihood that the entire jury will reach the correct conclusion in the case of ‘m’ voters chosen by ensemble voting and the probability ‘p’ taken into account for each voter to reach the correct conclusion is given as follows:

$$p_m = \sum_{i=\lceil m/2 \rceil}^{m} \left(\frac{m!}{(m-i)! \cdot i!}\right) \cdot p^i \cdot (1-p)^{m-1} (9)$$

For ex: if, $p > 5.0$, then $p_m > p$ i.e., in comparison to any individual voter there is a very high probability of accurate decision while using an ensemble voting.Figure. 1 shows the flowchart of the methodology that has been adopted in this work.

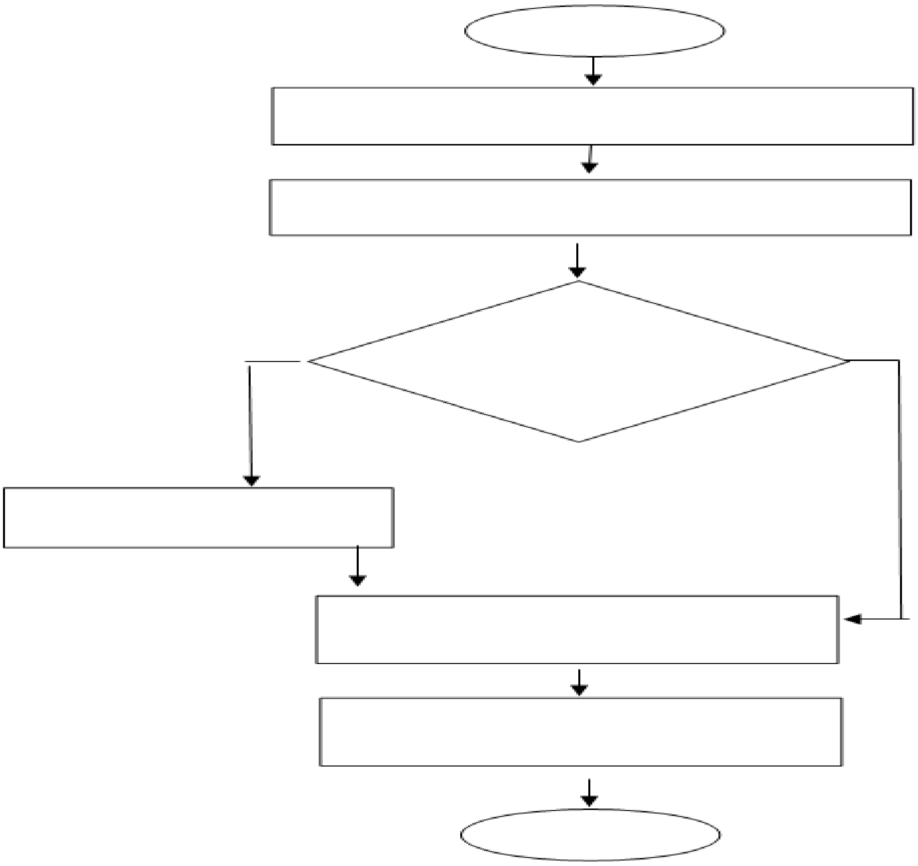

**Fig. 1: Proposed Methodology**

## RESULTS AND DISCUSSIONS

The proposed model presents the analysis of the COVID-19 time series data. For this, the dataset from Kaggle has been used.

**Table 1: Performance Analysis**

| Parameter | Gradient Boosting [21] | LinearRegression [21] | Xgboost Regression [22] | Proposed |
|---|---|---|---|---|
| MSE | 0.010 | 0.018 | 0.087 | 0.06 |
| RMSE | 1.32 | 0.18 | 0.29 | 0.13 |

The presented model has also been compared with models like Gradient Boositing, Linear Regression and XGboost regression in terms of parameters like MSE and RMSE. A shown in Table 1, the proposed model gives the best results. Table 2 shows the comparison of the proposed work with that of Prophet, Random Forest, XGboost for a different dataset.

**Table 2: Performance Analysis**

| Parameter | Prophet | Random Forest | XGBoost | Proposed Model |
|---|---|---|---|---|
| MAE | 3378.45 | 3668.06 | 3624.83 | 3218.17 |
| MSE | 17918614 | 17219729 | 16868326 | 15675791 |
| RMSE | 4233 | 4149 | 4107 | 4045 |

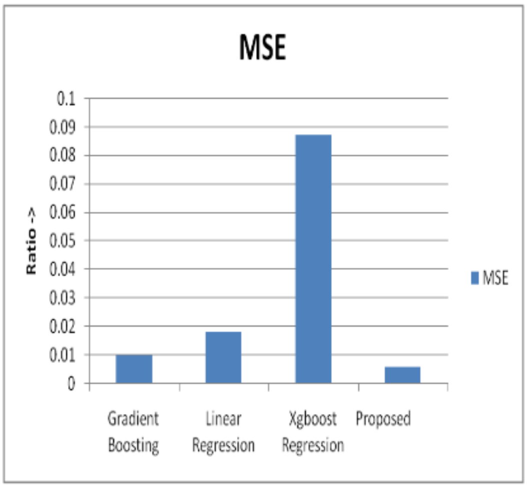

**Fig. 2: MSE analysis**

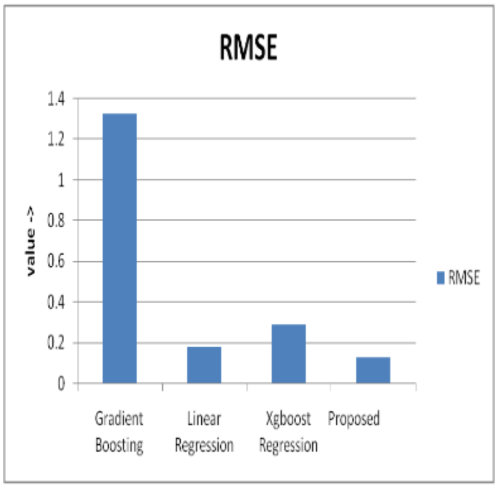

**Fig. 3: RMSE analysis**

Figure 2 and Figure 3, shows the comparative analysis of the proposed model with that of Gradient Boosting, Linear Regression, XGBoost respectively. The values of MSE and RMSE are compared. For both cases the proposed model has lesser values of error.

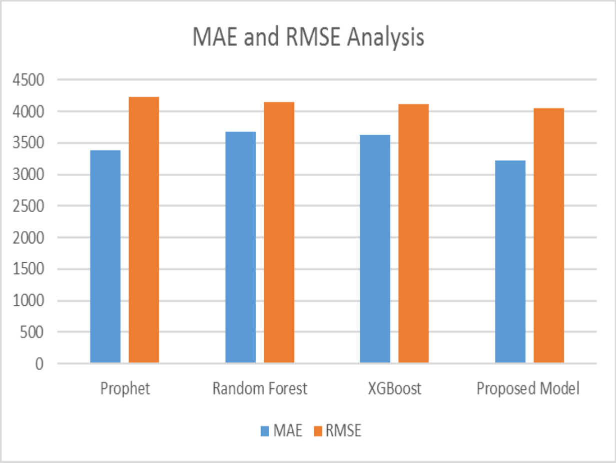

**Fig. 6 : MAE and RMSE Analysis**

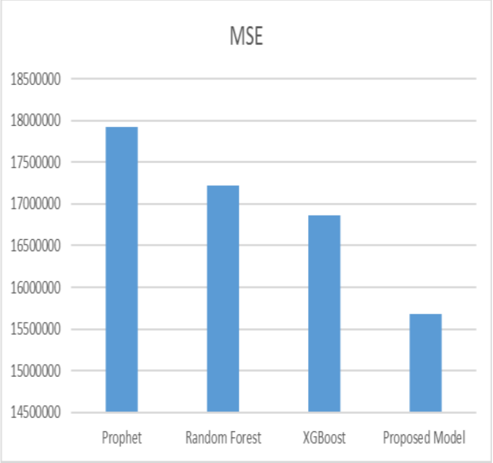

**Fig. 7: MSE analysis**

Figure 6 illustrates a comparison of the MAE and RMSE values of the various regression models for COVID-19 prediction. The MAE and RMSE values of the proposed model are the lowest. Figure 7 depicts the performance analysis of the proposed model in comparison with Prophet, Random Forest and XGboost.

## CONCLUSION

The motivation of the proposed method is clearly to aim for an accurate prediction of Covid-19. A time series-based data set consisting of confirmed and recovered cases and total deaths; from Kaggle has been used for this. A hybrid voting regression model has been proposed in the present. Python is used to implement the suggested model and MSE, MAE and RMSE have been calculated to support the model's performance. Analysis shows that the value of these parameters for the proposed models are lower than those of existing models. The results reinforce the use of the proposed model for COVID. The value of MAE for Prophet, Random Forest, XGBoost, and the proposed model is 3378.4, 3668.06, 3624.83 and 3218.17 respectively. The value of MSE for Prophet, Random Forest, XGBoost, and the proposed model is 17918614, 1719729, 16868326 and 15675791 respectively. The value of RMSE for Prophet, Random Forest, XGBoost, and the proposed model is 4233, 4149, 4107 and 4045 respectively.

## REFERENCES

1. Saud Shaikh, Jaini Gala, Aishita Jain, Sunny Advani, Sagar Jaidhara, Mani Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting", 2021, 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)
2. Arif Kurniawan, Fachrul Kurniawan, "Time Series Forecasting for the Spread of Covid-19 in Indonesia Using Curve Fitting", 2021, 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)
3. Mogens Graf Plessen, "Integrated Time Series Summarization and Prediction Algorithm and its Application to COVID-19 Data Mining", 2020, IEEE International Conference on Big Data (Big Data)

4. Suraj Bodapati, Harika Bandarupally, M Trupthi, “COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks”, 2020, IEEE 5th International Conference on Computing Communication and Automation (ICCCA)
5. Naresh Kumar, Seba Susan, “COVID-19 Pandemic Prediction using Time Series Forecasting Models”, 2020, 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)
6. Mohsen Mousavi, Rohit Salgotra, Damien Holloway, Amir H. Gandomi, “COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features”, 2020, IEEE Computational Intelligence Magazine, Volume: 15, Issue: 4
7. VenkatbharatPoleneni, Jahnavi K Rao, Syed Afshana Hidayathulla, “COVID-19 Prediction using ARIMA Model”, 2021, 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)
8. Abdul Jalil Niazai, Abdullah Zahirzada, Mohammad Akbar Shahpoor, Abdul Rahman Safi, “Time Series Forecasting of Registered, Recovered, and Death Cases of COVID-19 for the Next Sixty Days in Afghanistan”, 2020, IEEE International Conference on Advent Trends in Multidisciplinary Research and
9. Tingzhen Liu, Tong Zhou, Jin Gao, Wei Li, Yimin Ma, “Autocorrelation Sequence Prediction Model Based on Reference Function Transformation: Taking Epidemic Prediction As An Example”, 2020, Chinese Automation Congress (CAC)
10. Sina Ardabili, Amir Mosavi, Shahab S. Band, Annamaria R. Varkonyi-Koczy, “Coronavirus Disease (COVID-19) Global Prediction Using Hybrid Artificial Intelligence Method of ANN Trained with Grey Wolf Optimizer”, 2020, IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)
11. Raghavendra Kumar, Anjali Jain, Arun Kumar Tripathi, Shaifali Tyagi, “COVID-19 Outbreak: An Epidemic Analysis using Time Series Prediction Model”, 2021, 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)
12. Andi Sulasikin, YudhistiraNugraha, Juan Kanggrawan, Alex L. Suherman, “Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta”, 2020, IEEE International Smart Cities Conference (ISC2)
13. Zehua Yu, Xianwei Zheng, Zhulun Yang, Bowen Lu, Xutao Li, Maxian Fu, “Interaction-Temporal GCN: A Hybrid Deep Framework for Covid-19 Pandemic Analysis”, 20021, IEEE Open Journal of Engineering in Medicine and Biology
14. Radu Beche, Romina Baila, Anca Marginean, “COVID-19 spread forecast using recurrent auto-encoders”, 2020, IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)
15. Leonardo Sestrem de Oliveira, Sarah Beatriz Gruetzmacher, João Paulo Teixeira, “COVID-19 Time Series Prediction”, 2021, Procedia Computer Science
16. Suraj Bodapati, Harika Bandarupally, M Trupthi, “COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks”, 2020, IEEE 5th International Conference on Computing Communication and Automation (ICCCA)
17. Hanqing Chao, Xi Fang, Pingkun Yan, “Integrative analysis for COVID-19 patient outcome prediction”, 2020, Medical Image Analysis
18. Danish Rafiq, Suhail Ahmad Suhail, Mohammad Abid Bazaz, “Evaluation and prediction of COVID-19 in India: A case study of worst hit states”, 2020, Chaos, Solitons & Fractals
19. Farah Shahid, Aneela Zameer, Muhammad Muneeb, “Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM”, 2020, Chaos, Solitons & Fractals
20. Massimo A. Achterberg, Bastian Prasse, Piet Van Mieghem, “Comparing the accuracy of several network-based COVID-19 prediction algorithms”, 2020, International Journal of Forecasting Available online
21. Sina Ardabili, Amir Mosavi, Shahab S. Band, Annamaria R. Varkonyi-Koczy, “Coronavirus Disease (COVID-19) Global Prediction Using Hybrid Artificial Intelligence Method of ANN Trained with Grey Wolf Optimizer”, 2020, IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)
22. Zehua Yu, Xianwei Zheng, Zhulun Yang, Bowen Lu, Xutao Li, Maxian Fu, “Interaction-Temporal GCN: A Hybrid Deep Framework for Covid-19 Pandemic Analysis”, 20021, IEEE Open Journal of Engineering in Medicine and Biology.