Available online at http://scik.org
J. Math. Comput. Sci. 11 (2021), No. 3, 3647-3662
https://doi.org/10.28919/jmcs/5775
ISSN: 1927-5307

PREDICTIVE ANALYSIS OF NOVEL CORONAVIRUS USING MACHINE LEARNING MODEL - A GRAPH MINING APPROACH

PANKAJ KUMAR[1], RENUKA SHARMA[1], SURYA DEO CHOUDHARY[2,*], S. K. SINGH[3], VINOD M. KAPSE[2]

[1]Department of Computer Science and Engineering, Noida Institute of Engineering and Technology, Greater Noida - 201306, India

[2]Department of Electronics and Communication Engineering, Noida Institute of Engineering and Technology, Greater Noida - 201306, India

[3]Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida - 201306, India

Abstract: Graph mining is an important research field and it has received extensive attention. We can process, analyze, and extract meaningful information from a large amount of graph data. There are a large number of applications e.g. Biological network or web data. Machine learning is an example of biological networks algorithm. Machine learning and computational intelligence have promoted the development of predictive systems in a wide range of fields. These recommendations are based on contextual information that is explicitly provided or widely collected. Predictive systems usually improve solutions and increase task efficiency. Real-time data/information is not only a popular predictive system but also an abstraction of many real-world applications designed to increase resources and reduce risks. We can learn to use predictive analytics to predict the positive outcomes of these risks. These predictive analytics can look at the risks of past successes and failures. This paper attempts to develop an accurate (i.e. real-time) prediction recommendation system to predict new estimates of positive cases of coronavirus. Graph mining tool (i.e. machine learning) applied on the Indian dataset to predict the number of positive cases in daily, weekly, and monthly cases.

*Corresponding author
E-mail address: suryadeo.bit@gmail.com
Received March 27, 2021

Keywords: machine learning; COVID-19; computational intelligence; biological network; coronavirus; graph mining.

2010 AMS Subject Classification: 93A30.

1. INTRODUCTION

The COVID-19 pandemic is having a serious effect on the lives of billions of people throughout the world. Despite extensive safeguards such as national lockdown, international flight closures, and rigorous inspections, the spread of the infection continues to spread steadily, affecting thousands of people's death and has a serious socio-economic crisis. Unluckily, India is currently reporting the majority of cases in Asia. However, India's mortality rate is relatively low compared to world levels. Therefore, it is important to predict the number behind these infectious dynamics to keep down the effect and longevity of COVID-19 and future pandemics [1].

Technology is constantly evolving, and this process spreads rapidly over time. This development has solved many problems, but it has also brought new ones. One of these issues is the amount of data generated. Many sources can generate large amounts of data, such as medical records, mobile applications, automatic data creation, and customer databases. Due to the spread of the new coronavirus (SARS-CoV-2), the world is at a critical stage. So far, more than 5 million people have been affected. Around the world, people are very sensitive to future consequences. Early detection of positive cases is essential to prevent the spread of this epidemic and to treat patients quickly.

The novel coronavirus 2019 (COVID-19) is affecting 216 countries and their regions around the world. The cause is Severe Acute Respiratory Syndrome-Coronavirus 2 (SARS-CoV-2). The COVID-2019 first appeared in the city of China, Wuhan on 31 December 2019, and then speedily spread throughout the globe and became a pandemic. It has had a terrifying impact on day-to-day life, public health, and the worldwide economy. It is crucial to find out the positive cases as early as possible to prevent the epidemic from spreading further and treat patients immediately. The virus can infect people, kill people, infect others, or simply cure the disease. The vaccination and the immune system can often fight illnesses and help people reduce the effects of illnesses that are still infectious. The lack of accurate automation tools increases the demand for diagnostic tools under certain circumstances. Decision-making analysis and prediction of positive case data show that graph mining in the medical field is very promising.

Graph mining is the study of how to perform data mining and machine learning on graph data. It is a set of tools and techniques used to analyze graph properties in real-time, can predict how a particular graph structure and properties affect a particular application and develop a model that can generate the actual graph, which matches the model found in real life [2]. Graph mining techniques has categorized as follows [3].

- *Graph Clustering*: An input graph is divided into multiple clusters using their vertices. It is based on unsupervised learning where the class is unknown before clustering. The cluster of graphs is formed based on the similarities of the main structured data graph.
- *Graph Classification*: The main task in classifying statistical graphs is to divide the individual statistical graphs and individual statistical graphs in the statistical graph database into two or more categories/classes. This classification is based on supervised learning / semi-supervised learning with predefined data categories.
- *Subgraph Mining*: A subgraph is a subset of the vertices and edges of a single graph. A common problem when mining subgraphs is to create a set of subgraphs that are at least displayed at a particular threshold for a particular n-sample input graph.

Machine learning based on various types of applications and research conducted by industry and academia is changing the world rapidly. It affects every aspect of our daily lives. It behaves like the growth of a child. As children grow older, they also have more experience completing tasks, which leads to a higher performance measure. Machine learning models help us perform many tasks such as object recognition, synthesis, prediction, classification, and clustering. Therefore, this paper uses machine-learning algorithm as data analysis and prediction tool.

The paper is organized as follows: Section 2 illustrates the related works. Section 3 proposes the methodology. Section 4 describes the materials and methods. Section 5 describes the dataset training and testing. Section 6 describes the results, analysis, and discussion. Finally, Section 7 ends with conclusion.

2. RELATED WORKS

Designing an accurate prediction system in a real application can be difficult. Modeling a real-time series is especially difficult because it usually consists of linear and non-linear models that are somehow combined [4, 5]. To solve the prediction problem, there are several creature-inspired metaheuristics networks. Due to the difficulty of consensus among researchers on which method to apply to which problem, several predictive methods have been proposed in the last decade to improve machine learning models. In general, from a detailed point of view, the criteria for choosing a method are related characteristics. Computational cost, accuracy, and even reduction of implementation complexity can be considered one of these criteria [18].

Since the advent of COVID-19, which has spread to the continents of developed and developing countries, several research papers have been published on various aspects of COVID-19. Several machine learning models have been developed to estimate the number of positive cases for COVID-19. Using certain, types of algorithms and certain restricted parameters constrain the model. A variety of methods and techniques can be used to predict the number of COVID-19-positive cases. Machine learning makes it easy to reuse past and existing data and save new lives. However, researchers and practitioners use machine-learning algorithms to make the best predictions [6, 7].

The Table 1 summarizes a series of baseline studies related to the use of machine learning to predict COVID-19-positive cases.

Table 1: Related Works

Year	Paper Title and Authors Name	Objectives
2020	Automated detection of COVID-19 cases using deep neural networks with X-ray images Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim,U. Rajendra Acharya	Launch of a new model that automatically detects COVID-19 using the original chest X-ray. The proposed model is designed to provide accurate diagnosis with binary classification (no results) (COVID vs. no results) and binary classification (COVID vs. no results vs. pneumonia). [12]
2020	A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2 Mohammad Rahimzadeh, Abolfazl Attar	The author introduced several complex and deep networks through the introduction of training methods. The training method is based on two open source datasets and divides x-rays into three categories: normal, pneumonia, and COVID-19. [13]
2020	Prediction for the spread of COVID-19 in India and effectiveness of preventive measures Anuradha Tomar, Neeraj Gupta	The authors used data-driven scoring methods (such as long short-term memory and curve fitting) to describe the number of COVID-19 cases in India 30 days ago, with social isolation and COVID- 19 Prevented epidemics. Authors predict the outcome of precautionary measures. Prediction is done using various parameters (number of positive cases, number of cured cases, etc.). [14]

2020	Coronavirus Optimization Algorithm: A Bioinspired Metaheuristic Based on the COVID-19 Propagation Model F. Martı´nez-Alvarez, G. Asencio-Corte´s, J. F. Torres, D. Gutie´rrez-Avile´s, L. Melgar-Garcıa, R. Pe´rez-Chaco´ n, C. Rubio-Escudero, J. C. Riquelme, and A. Troncoso	The authors propose a new metaheuristic method that can mimic the spread of coronavirus and infection in healthy people. Starting with the main infected (zero patients), the coronavirus quickly infects new victims, killing many or spreading the infection. In order to simulate the activity of the coronavirus as much as possible, we introduce relevant terms (potential reinfection, hyperactive diffusion rate, social distance measurement, speed of travel, etc.) into the model. [11]
2020	Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India Parul Arora, Himanshu Kumar, Bijaya Ketan Panigrahi	An intensive training model (i.e. deep learning model) is Apply to estimate the number of new cases of coronavirus (COVID-19) infection in 32 Indian states and territories. Long short-term memory (LSTM) variants (deep LSTM, convolutional LSTM, bidirectional LSTM, etc.) based on recurrent neural networks (RNNs), are applied to Indian datasets to predict the number of positive cases. [6]
2020	A data-driven understanding of COVID-19 dynamics using sequential genetic algorithm based probabilistic cellular automata Sayantari Ghosh, Saumik Bhattacharya	The authors believe that for this accurate data-driven modeling of infection transmission, cellular automata provide an excellent sequential gene algorithm platform that can effectively estimate dynamic parameters. [15]
2020	Analysing the Covid-19 Cases in Kerala: a Visual Exploratory Data Analysis Approach Jayesh S, Shilpa Sreedharan	In this paper, authors will analyze the case of Covid-19 and take effective action by the Indian government of Kerala in response to the outbreak of Kerala. [16]

3. PROPOSED METHODOLOGY

Machine Learning research is currently undergoing major changes to gain an accurate view of the future. The term is known as time series prediction when a mathematical model predicts that future data will be recorded only in terms of time [9]. There are many machine learning and deep learning algorithms that can be used for time-series prediction, such as ARIMA and LSTM. This paper describes a machine learning algorithm i.e. ARIMA to predict the number of suffered, and recovered patients in India. Figure 1 shows the proposed methodology.

Boxing and Jenkins have proposed the ARIMA model. It is an abbreviation for AutoRegressive Integrated Moving Average. It is one of the simplest and most efficient machine learning algorithms for time series prediction. This is a combination of automatic regression and moving average [17].

This summary is descriptive and covers important aspects of the model itself. These include:

- *AutoRegression* (AR): This is a time series model that uses the observations from the previous time step as input to the regression equation and predicts the values in the next step. The final time step t-1 is used to predict t.

  $$Y_t = \alpha_1 Y_{t-1} + \varepsilon_t \dots\dots\dots\dots (1)$$

  Y_t is a linear function of the previous value and is defined in Equation 1. Each observation has a random component (ε) and a linear combination of previous observations. This equation has an autoregressive coefficient α_1.
- *Integrated* (I): An integrated process that uses interpolation from raw observations to keep the time series constant (for example, subtracting observations from observations from the previous step) is given in Equation 2.

  $$Y_t = Y_{t-1} + \varepsilon_t \dots\dots\dots\dots (2)$$

  ε_t is the noise.
- *Moving Average* (MA): It is also known as the rolling mean. Calculate a simple average for a period and divide it by the total number of periods obtained. Moving average is defined by equation 3.

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} \dots\dots\dots\dots (3)$$

Any from these elements is explicitly specified as a parameter in the ARIMA model. ARIMA uses standard notations (p, d, q) and replaces parameters with integer values to indicate that a particular model is being used quickly. The parameters of ARIMA are defined as follows:

- p: it is he number of interval observations included in the model also known as an interval sequence.
- d: it is the frequency with which raw observations are distinguished is also called the degree of distinction.
- q: it is the size of the moving average window also known as a moving mean.

The linear regression model is formed with a specified amount and variety of terms, and the data is modeled to varying degrees, eliminating static and seasonal configurations that make the regression model negative. It affects the shape.

We can use the value 0 (Zero) for the parameter. It indicates that this model element should not be used. Therefore, the ARIMA model can be tuned to perform the functions of the ARMA model and even simple AR, I, or MA models.

After selecting the ARIMA time series model, the underlying process for generating observations is assumed to be the ARIMA process. It may sound obvious, but raw observations trigger the requirement to test model hypotheses and residual errors in model predictions. The time series can be divided into three elements.

- Trends: Up and down movements over a long period
- Season: Seasonal variation
- Noise: Spikes & troughs at random intervals

ARIMA model uses brute force to experiment with different combinations of p, q, and d to get the best model after evaluation. Estimate the optimal model using the root mean square error. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which are statistical measures of model fit quality and simplicity, are also used.

4. MATERIALS AND METHODS

a) Machine Learning

Machine learning is a subset of the scientific field of artificial intelligence, focusing on the study of computer algorithms that can automatically improve computer programs based on experience and make predictions based on experience [8. 10]. Use the training dataset to train machine-learning algorithms (i.e. ARIMA) to build models. When new input is entered into the machine-learning algorithm, predictions are made based on the model. Evaluate the accuracy of the predictions and, if the accuracy is acceptable, deploy the machine-learning algorithm. If accuracy is unacceptable, use the extended training dataset to retrain the machine-learning algorithm.

The fact that the number of COVID cases is increasing rapidly does not mean that the data can be superimposed on the exponential curve to estimate the number of cases in the coming days. Compartmental modeling technology is typically used to model infectious diseases. It can also be used for COVID-19. Therefore, we need a reliable model for predicting how the virus will spread in different parts of India. We chose machine learning, a graph mining tool because the purpose of this task is to build a model that predicts the spread of the virus. To predict infectious diseases and mortality in India, we propose an intuitive COVID-19 prediction model that adds machine learning techniques to the classic infectious disease model.

b) Implementation in Python Programming

Python prediction focuses on the skills needed to master the context of the entire problem. We always keep the decision in mind and make the best business decisions. The Python programming language is well suited for data science, machine learning, system automation, web and API development, and more. Python is a high-level, descriptive object-oriented programming language with dynamic semantics. The high-level integrated data structure that combines dynamic typing and dynamic linking makes it very attractive as a scripting or pasting language for rapid application development and linking existing components.

Over the years, Python has grown into a leader in modern software development, infrastructure management, and data analytics. It is no longer a service language for office space; it is the main engine for building web applications and systems management, and for the explosive growth of big data analytics, and artificial intelligence. We can analyze and predict time series data using open source libraries such as Numpy, Pandas, matplotlib, prophet, seaborn, and datetime. Python is a high-level programming language used to interact with machine learning libraries such as application programming interfaces (APIs).

c) Dataset Description

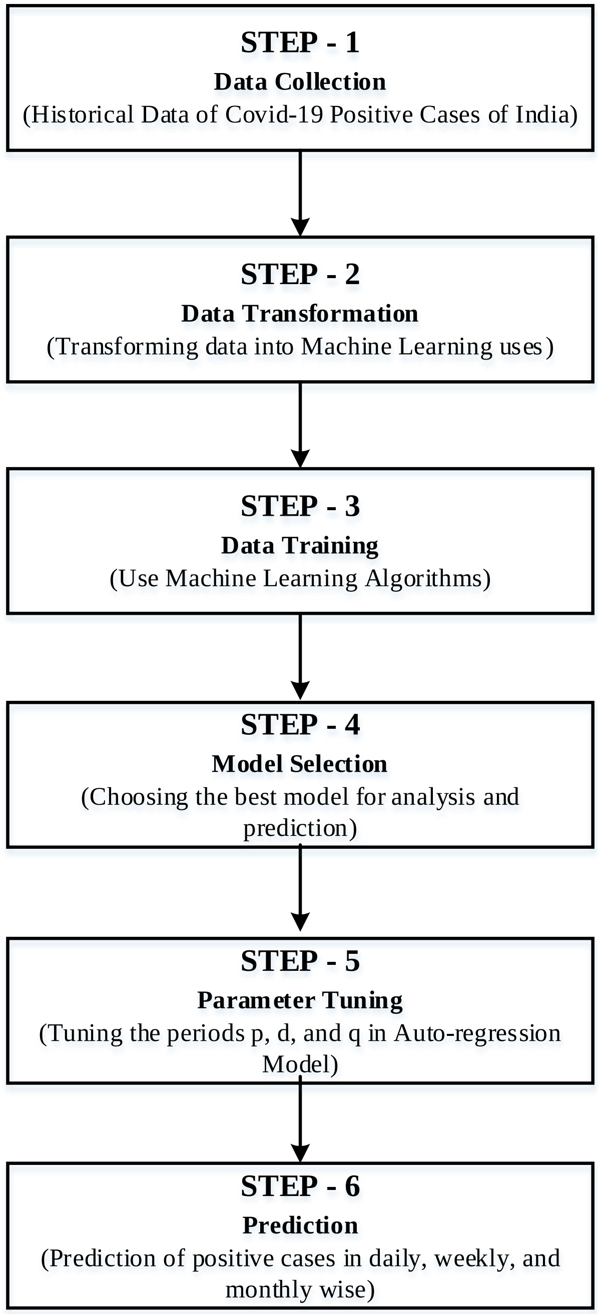
There are many formal and informal dataset sources on the Internet, including data related to COVID-19. The dataset was obtained from Worldometer and the Ministry of Health and Family Welfare (Government of India). These numbers are inherently very stable, as the increase or decrease in incidence depends on other environmental / physical factors. They include the duration of COVID-19 cases identified in each state (28) and Union Region (8) from 30 January 2020.

Use the COVID-19 dataset to understand how the disease has spread to different parts of India. In addition to datasets, it also performs data processing and visualization operations. We can also use a linear regression algorithm (i.e. ARIMA) to understand the number of valid active and recovered cases.

5. DATASET TRAINING AND TESTING

We are analyzing to predict cases of COVID-19 available at https://www.worldometer.info/coronavirus/country/india/ [1]. Initially, the dataset is collected from January 30, 2020, to October 31, 2020. 80% of the data will be used for training, and the remaining 20% will be used for predictive and validation purposes. For training and testing of our data, ARIMA machine-learning model via Python is used.

Figure 1: Layout of the Proposed Methodology



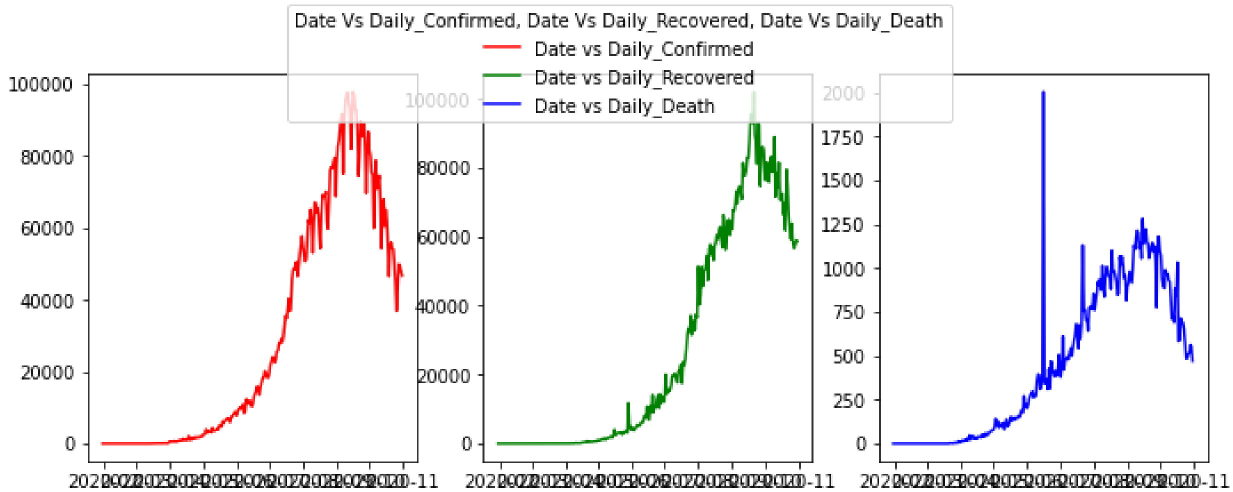


Figure 2: Date vs. Daily confirmed, Daily death, and Daily recovered

6. Results, Analysis, and Discussion

Visualization of raw data is easy to understand always. Here we compare the evolution of Covid-19, which has been confirmed dead, with a severely infected case in India. Visualizations are created in Python using the matplotlib, prophet, seaborn, arima, and datetime libraries to analyze time series data. Figure 2 shows the variation of the daily covid-19 cases as confirmed, recovered, and death. The cases has so much fluctuation. As one variable increases, so does the other. Charts can be used to describe different Covid-19 cases based on events over time. Each graph shows the data and its changes from January 30 to October 31. Here, the first graph shows the date and total number of confirmed cases, the second graph shows the date and total number of deaths, and the third graph shows the date and total number of cured cases.

Figure 3 shows the output of all possible combinations of attributes in the dataset as a scatterplot. We can see the correlation between each column / attribute to another column / attribute in the graph of the Covid-19 dataset. Use Seaborne as a multidimensional visualization tool to show the correlation between columns as scatterplot. In this scatterplot, the diagonal cells show histograms of each of the columns, in this case the concentrations of the six columns (Total No. of Cases, Day-wise New Cases, Total No. of Recovered, Day-wise New Recovered, Total No. of Death, and Day-wise New Death). Each of the off-diagonal cells is a scatterplot of two of the six columns. The pattern of data points is different. They are not randomly distributed on the diagram.

We can draw a diagonal line next to most data points.

The six pairs of attributes are correlated. Correlation matrices that define linear relationships between different characteristics are not a good sign, as shown in Figure 3 and Figure 4. As one variable increases, so does the other.

A heatmap is a two-dimensional graphic representation of dataset, in which the individual values contained in the matrix are displayed in color. The Seaborn Python package allows us to create annotated heatmap of real-time data that we can edit as needed using the Matplotlib tool. We can create a Python Seaborn heatmap of any size. The multiple attributes classification performance of the ARIMA machine learning model has been evaluated for each fold, and the average classification performance of the model is calculated. The multiple attributes classification heatmap is shown in Figure 5.

Each square (Figure 5) shows the correlation between the variables on each axis in Figure 4. The range of correlation is -1 to +1. A value close to zero means that there is no linear trend between the two attributes. Next to correlation 1, there is a strong positive correlation between the two attributes. This means that daily recovery and death have not improved. A correlation close to -1 is similar, but increasing one variable instead of both increases the other attributes. Analysis of the correlation matrix shows that the total number of cases (TC) has a strong positive correlation with the total number of deaths (DT) (e.g. 0.99) and the total number of recovered cases (TR) is showing 1. TC has a strong negative correlation with TR on the graph, but it is a good sign of recovery. Since these squares associate each variable with itself, the diagonals are all 1 (hence, this is a perfect correlation). The correlation between the two variables is higher for static numbers and darker colors. The graph is also diagonally symmetric because these squares sum the same two variables.

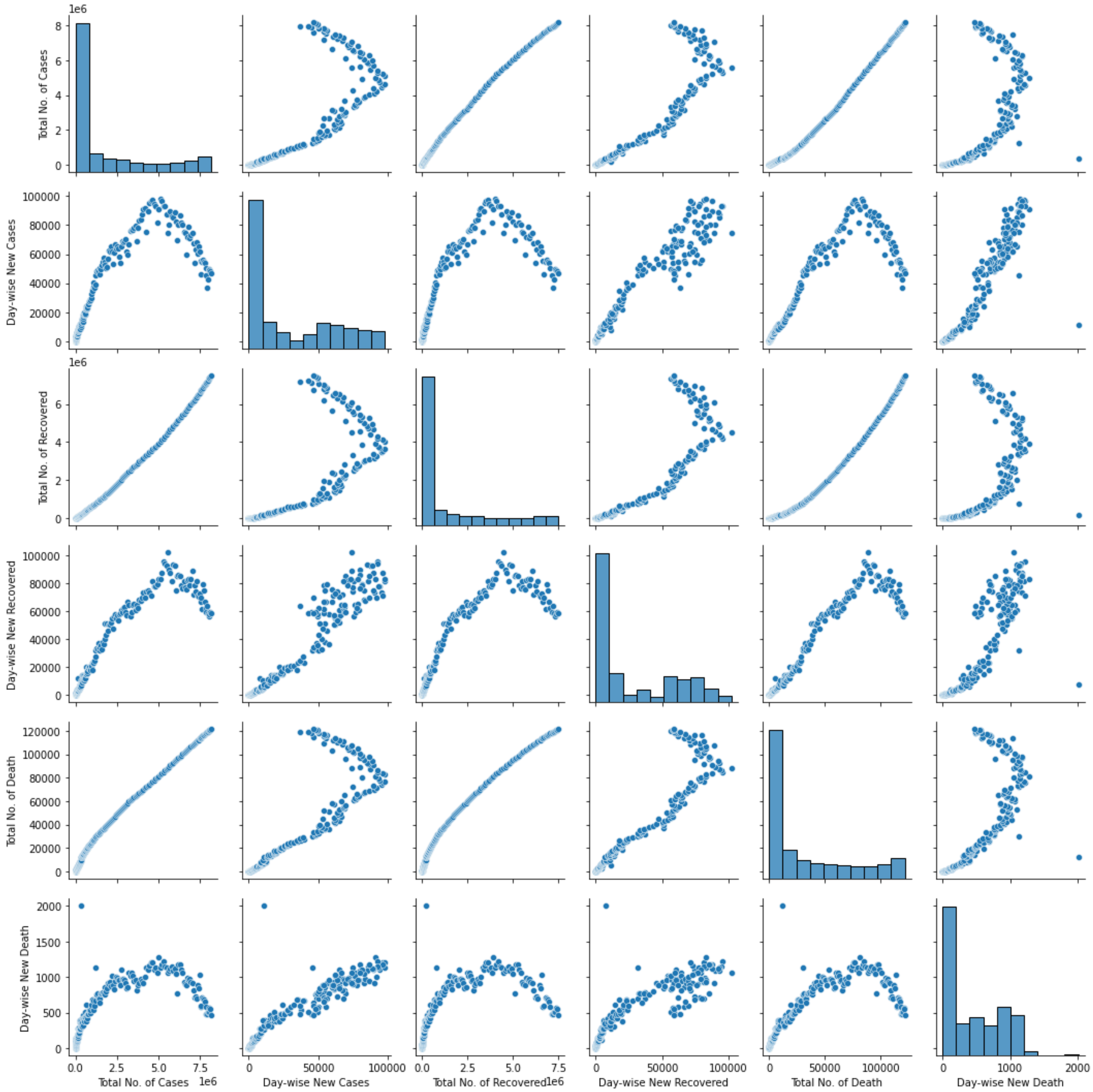


Figure 3: Multidimensional Correlation of Attributes as scatterplot

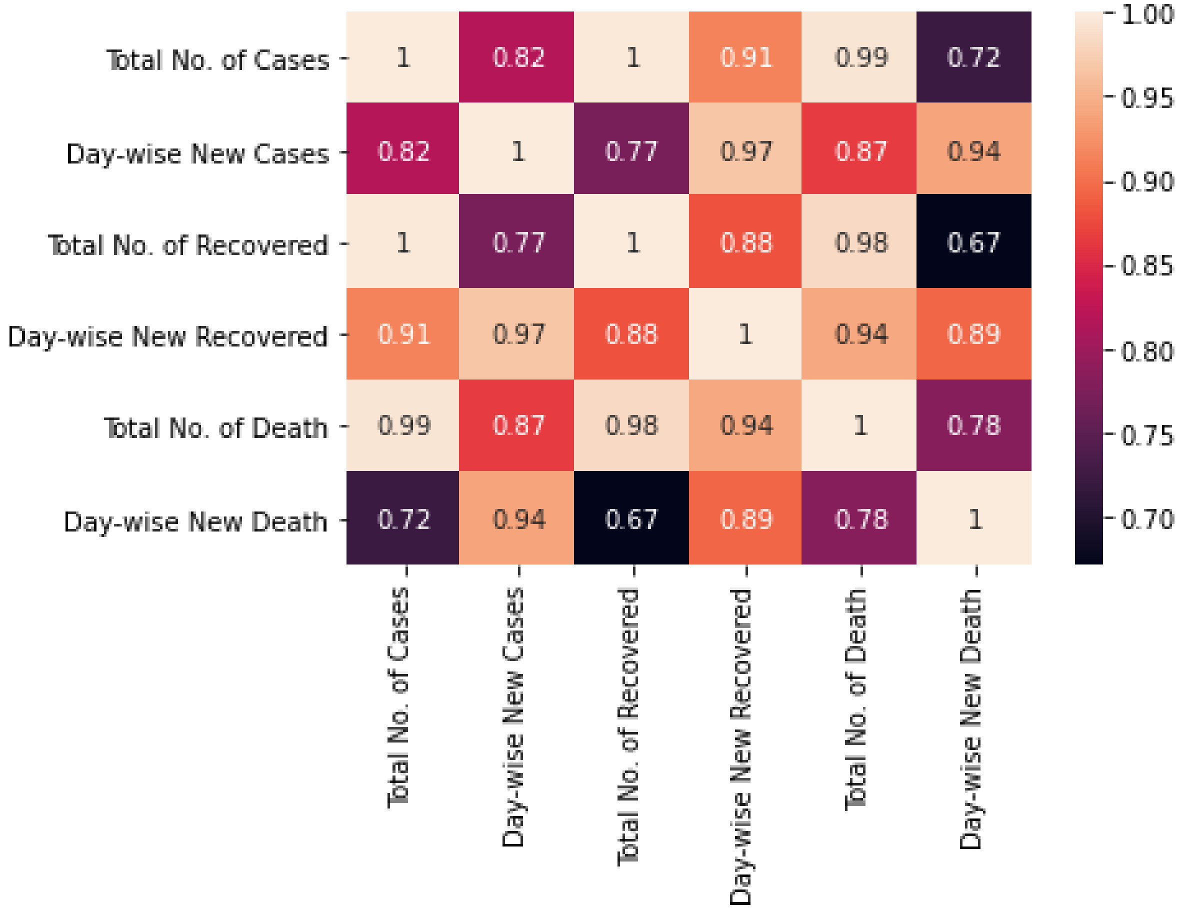


Figure 4: Correlation Matrix as a heatmap

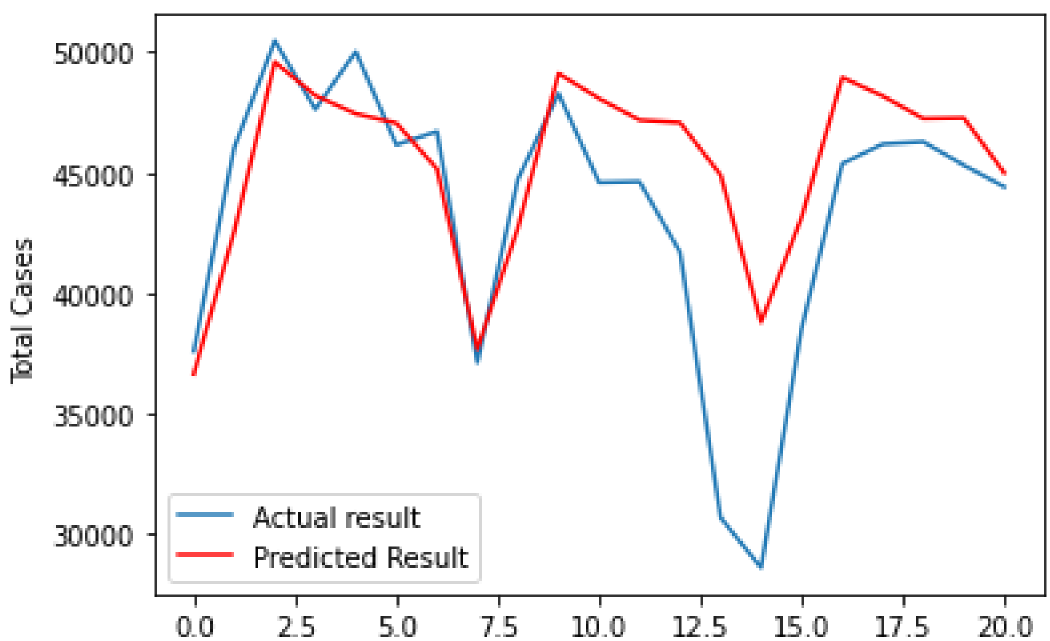


Figure 5: Graph of actual vs. predicted total no. of confirmed cases

Table 2: Predicted total no. of confirmed cases

S. NO.	Date	Actual Cases	Predicted Cases
0	02/11/2020	37592	36639.1100
1	03/11/2020	46027	42550.1069
2	04/11/2020	50465	49585.6778
3	05/11/2020	47628	48196.7176
4	06/11/2020	50017	47443.2300
5	07/11/2020	46153	47056.2746
6	08/11/2020	46707	45152.2959
7	09/11/2020	37119	37677.7126
8	10/11/2020	44724	42704.5735
9	11/11/2020	48285	49111.7714
10	12/11/2020	44585	48075.4681
11	13/11/2020	44620	47182.4335
12	14/11/2020	41692	47074.3080
13	15/11/2020	30681	44891.5208
14	16/11/2020	28609	38790.3697
15	17/11/2020	38548	43148.2006
16	18/11/2020	45366	48950.9872
17	19/11/2020	46185	48179.7341
18	20/11/2020	46283	47241.8143
19	21/11/2020	45301	47268.0768
20	22/11/2020	44404	44982.5902

7. CONCLUSION

This paper proposes a machine learning predictive model used to evaluate the possible number of confirmed COVID-19 cases in real-time, based on confirmed cases in India. The model we have developed can predict confirmed cases with the accuracy and efficiency shown in Figure 5 and Table 2. Divide the dataset into two parts. The first part uses 80% of the data for training purposes, and the second part uses 20% for prediction and validation purposes. The number of COVID-19 cases identified daily was estimated using the ARIMA model. The ARIMA model is tested on an

Indian dataset and the most accurate model is selected based on absolute error. Based on prediction errors, ARIMA gives the best results, and Facebook prophets give the worst results. Daily and weekly forecasts for India have been calculated and ARIMA has been found to provide very accurate results (less than 2% error) for short-term weekly forecasts. This model can be used in remote areas of the country that are more severely affected by COVID-19 to help overcome the shortage of tests and hospitals. We intend to make the model more reliable, accurate, and efficient to increase the number of datasets.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] Worldometer: https://www.worldometers.info/coronavirus/country/india/ [Last visited: 24/11/2020].

[2] S.U. Rehman, A.U. Khan, S. Fong, Graph mining: A survey of graph mining techniques, in: Seventh International Conference on Digital Information Management (ICDIM 2012), IEEE, Macau, Macao, 2012: pp. 88–92.

[3] S. Aridhi, E. Mephu Nguifo, Big graph mining: frameworks and techniques, Big Data Res. 6 (2016) 1–10.

[4] P. Kumar, S.K. Singh, A framework for assessing the evolvability characteristics along with sub-characteristics in AOSQ model using fuzzy logic tool, in: 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, Greater Noida, 2017: pp. 339–344.

[5] S.F. Ardabili, A. Mosavi, P. Ghamisi, et al. COVID-19 outbreak prediction with machine learning, Algorithms. 13 (2020), 249.

[6] P. Arora, H. Kumar, B.K. Panigrahi, Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India, Chaos Solitons Fractals. 139 (2020), 110017.

[7] Z. Car, S. Baressi Šegota, N. Anđelić, I. Lorencin, V. Mrzljak, Modeling the spread of COVID-19 infection using a multilayer perceptron, Comput. Math. Meth. Med. 2020 (2020), 5714714.

[8] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. 20 (2020), 533–534.

[9] P. Kumar, S. Dixit, S. K. Singh, Performance of aspect-oriented software quality modelling using artificial neural network technique, Int. J. Computer Appl. 182 (2019), 6–10.

[10] P. Kumar, S.K. Singh, An Emerging Approach to Intelligent Techniques—Soft Computing and Its Application, in: V.E. Balas, V.K. Solanki, R. Kumar (Eds.), Internet of Things and Big Data Applications, Springer

International Publishing, Cham, 2020: pp. 171–182.

[11] F. Martínez-Álvarez, G. Asencio-Cortés, J.F. Torres, et al. Coronavirus optimization algorithm: a bioinspired metaheuristic based on the COVID-19 propagation model, Big Data. 8 (2020), 308–322.

[12] T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U. Rajendra Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images, Computers Biol. Med. 121 (2020), 103792.

[13] M. Rahimzadeh, A. Attar, A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2, Inform. Med. Unlocked. 19 (2020), 100360.

[14] A. Tomar, N. Gupta, Prediction for the spread of COVID-19 in India and effectiveness of preventive measures, Sci. Total Environ. 728 (2020), 138762.

[15] S. Ghosh, S. Bhattacharya, A data-driven understanding of COVID-19 dynamics using sequential genetic algorithm based probabilistic cellular automata, Appl. Soft Comput. 96 (2020), 106692.

[16] J. S, S. Sreedharan, Analysing the Covid-19 cases in Kerala: a visual exploratory data analysis approach, SN Compr. Clin. Med. 2 (2020), 1337–1348.

[17] D.S. de O. Santos Júnior, J.F.L. de Oliveira, P.S.G. de Mattos Neto, An intelligent hybridization of ARIMA with machine learning models for time series forecasting, Knowl.-Based Syst. 175 (2019), 72–86.

[18] P. Kumar, S.K. Singh, A systematic assessment of aspect-oriented software development (AOSD) using JHotDraw application, in: 2016 International Conference on Computing, Communication and Automation (ICCCA), IEEE, Greater Noida, India, 2016: pp. 779–784.