Role of Artificial Intelligence in COVID-19 Prediction Based on Statistical Methods

5

R. Sujatha and Jyotir Moy Chatterjee

Abstract

Coronavirus disease 2019 (COVID-19) is a respiratory ailment that can spread from individual to individual, first recognized during an outbreak in Wuhan, China. Possibility of getting COVID-19 is greater for people who are in contact of someone known to have COVID-19, for example clinical professional, or family member. Peoples at higher peril for sickness are the those who live in or have starting late been in a zone with advancing spread of COVID-19. A few patients have pneumonia in the two lungs, multi-organ failure and sometimes death. In this work we are predicting the impact of COVID-19 cases in India based on time series analysis, correlation analysis, Granger Test, and Group Method of Data Handling (GMDH). We have compared the prediction of four algorithms namely combinatorial (quick), stepwise forward, stepwise mixed, and GMDH neural network (GMDH-NN) for predicting the future of India. Out of that stepwise mixed method gives good prediction for confirmed cases but GMDH-NN gives better prediction in case of death and recovered cases. As this disease is declared as an epidemic, the present study will help researchers to understand the impact of this outbreak. We have used combinatorial (quick), stepwise forward selection, stepwise mixed selection, and GMDH neural network to predict the spread of disease in India.

R. Sujatha (✉)
SITE, Vellore Institute of Technology, Vellore, India
e-mail: r.sujatha@vit.ac.in

J. M. Chatterjee
Lord Buddha Education Foundation, Kathmandu, Nepal

S. Nandan Mohanty et al. (eds.), *Applications of Artificial Intelligence in COVID-19*, Medical Virology: From Pathogenesis to Disease Control,
https://doi.org/10.1007/978-981-15-7317-0_5

Keywords

COVID-19 · Time series analysis · Combinatorial (quick) · Stepwise forward selection · Stepwise mixed selection · GMDH neural network (GMDH-NN)

5.1 Introduction

The viral infection that causes COVID-19 most likely rose out of a creature, but is currently spreading from individual to individual. Overall, COVID-19 is a challenge faced by multiple disciplines like medicine, defense, finance, telecommunication, information technology, and so on. The contamination is thought to spread chiefly between people who are in close contact with one another (inside around 6 ft) through respiratory globules conveyed when a infected individual hacks or sniffles. It also may be possible that an individual can get COVID-19 by reaching a surface or article that has the disease on it and a short time later reaching their own mouth, nose, or conceivably their eyes, yet this isn't accepted to be the essential way the contamination spreads. Patients with COVID-19 have had delicate to genuine respiratory malady with impacts of:

- Fever
- Cough
- Shortness of breath

The World Health Organization (WHO) announced the 2019–20 coronavirus epidemic a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 (WHO 2020; Mahtani 2020) and a pandemic on 11 March 2020 . Proof of neighborhood spread of the ailment has been found in numerous nations over each of the six WHO districts (World Health Organization 2020a, b).

5.2 Related Work

Xu et al. (2020) examined the neurotic characteristics of a patient who passed on from severe infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by postmortem biopsies.

Novel (2020) reported outcomes of a descriptive, experimental investigation of all cases identified as of February 11, 2020.

Chen et al. (2020) expected to assess the clinical characteristics of COVID-19 in pregnancy and the intrauterine vertical transmission capability of COVID-19 infection.

Liu et al. (2020) audit the basic reproduction number (R0) of the COVID-19 virus. R0 means that the transmissibility of a virus, representing the normal number of new infections created by an infectious person in an absolutely innocent populace.

Pan et al. (2020) studied the adjustment in chest CT findings associated with COVID-19 pneumonia from beginning diagnosis until understanding recuperation.

Sujatha et al. (2020a, b) utilized linear regression, vector autoregression and multilayer perceptron technique for COVID-2019 cases prediction in India using Kaggle dataset.

Iwendi et al. (2020) applied boosted random forest algorithm for COVID-2019 prediction.

5.3 Dataset Description

We have used the COVID-2019 (KP 2020) dataset from kaggle where data from January 22, 2020 to March 26, 2020 are present. The dataset is having data of more than 180 countries with their attributes such as Province/State, Country/Region, Latitude, Longitude, Date, Confirmed, and Deaths. Out of this dataset we have concentrated on India's data. The dataset is having 65 instances for India. As we have taken only India's data so with have discarded the latitude and longitude by only considering the corresponding dates followed by the confirmed, death, and recovered cases. We anticipated the future effects of COVID-19 pandemic in India through time series analysis, correlation analysis and Granger Test and GMDH.

5.4 Experimental Results

For experimentation purpose we have incorporated Group Method of Data Handling (GMDH) strategy. This method was started by Ivakhneko in 1966 and it has been improved and advanced in the course of recent years. The GMDH calculation interfaces the inputs to outputs with higher sequential polynomial networks which are principally feed-forward and multilayered neural networks (NN) (Onwubolu 2009). Right now, nodes are shrouded units and the activation polynomial coefficients are weights which are evaluated by standard least square regression (Ghanadzadeh et al. 2012). Lately, be that as it may, the utilization of such self-composed networks has prompted fruitful use of the GMDH-type computation in a wide scope of zones in engineering and science (Ahmadi et al. 2007; Abdolrahimi et al. 2014; Pazuki and Kakhki 2013; Atashrouz et al. 2015; Najafzadeh 2015). The GMDH is a polynomial-based model. As indicated by the GMDH approach, each layer can be acquired from a quadratic polynomial function. In this manner, the input variables are anticipated to the yield variable. The primary objective right now is finding of function, *f*, which ventures the input variables to the yield variable. In this way, the output variable (X_i) can be composed from the input variables as the accompanying structure:

$$X_i = f(Y_{i1}, Y_{i2}, Y_{i3} \ldots Y_{in}), \quad i = (1, 2, 3, \ldots, M) \tag{5.1}$$

where, Y_s are input variables. The structure of the GMDH can be obtained using the minimization of an objective function. The objective function ω can be written as:

$$\omega = \sum_{i=1}^{M} [(\mathrm{X}(Y_{i1}, Y_{i2}, Y_{i3} \ldots Y_{in})) - X_i]^2 \tag{5.2}$$

where, in the above equation X_i is actual data (Naderpour et al. 2019).

GMDH (Ivakhnenko 1971; Dorn et al. 2012) is known as a self-sorting deep learning technique for time-series analysis issues. It is broadly utilized in numerous fields, for example, forecasting, data mining, optimization and pattern recognition and so on. GMDH-based NN can be considered as a polynomial NN. As a distinction with different networks, the GMDH network changes persistently during the preparation procedure. A few favorable circumstances of GMDH network can be referenced as: self-association in the preparation procedure, high exactness in forecasting, findings for high-request nonlinear frameworks and so forth (Nguyen et al. 2019).

GMDH comprises of parametric, clusterization, analogs complexing, re-binarization and likelihood calculations. This inductive methodology depends on sifting through of bit by bit entangled models and determination of the ideal arrangement by least of outer standard trademark. Polynomials as well as nonlinear, probabilistic methods or clusterizations are utilized as essential models.

GMDH methods can be valuable in light of the fact that:

- The number of layers and neurons in hidden layers, model structure and other ideal hyperparameters are resolved consequently.
- It ensures that the most precise or impartial models will be found—technique doesn't miss the best arrangement during arranging everything being equal (in the given class of functions).
- As input variables can be utilized nonlinear functions, that may have effect on output variable.
- It naturally finds interpretable connections in data and chooses compelling input factors.
- GMDH sorting calculations are fairly basic for developing software (GMDH n.d.) .

We have used the GMDH software and Orange Data Mining software (Bioinformatics Laboratory 2020) for conducting our experiments. Granger test and correlation analysis is conducted using and time series analysis is conducted using (NNS 2020). Data science begins with collecting the data followed by consolidating, investigating, understanding and finally presenting it with valuable information. In data consolidating and investigating phase, as a part of knowledge gain, should know the nature of attributes and the nature of values make up the dataset. Data visualization helps in getting great insight of the data set followed by applying the required process of classification, association, clustering based on the problem scope

Table 5.1 Statistics of INDIA COVID-19

Variable	Date	Confirmed	Deaths	Recovered
Numeric values	0	65	65	65
Text values	65	0	0	0
Missing values	0	0	0	0
Unique values	65	28	10	13
Zero values	0	8	49	25
Most frequent		3	0	0
Min. value		0	0	0
Max. value		727	20	45
Median		3	0	3
Mean value		73.72	1.35	6
Std. deviation		157.60	3.43	10.52
2σ outliers		5	4	4
3σ outliers		2	2	3
4σ outliers		1	1	0

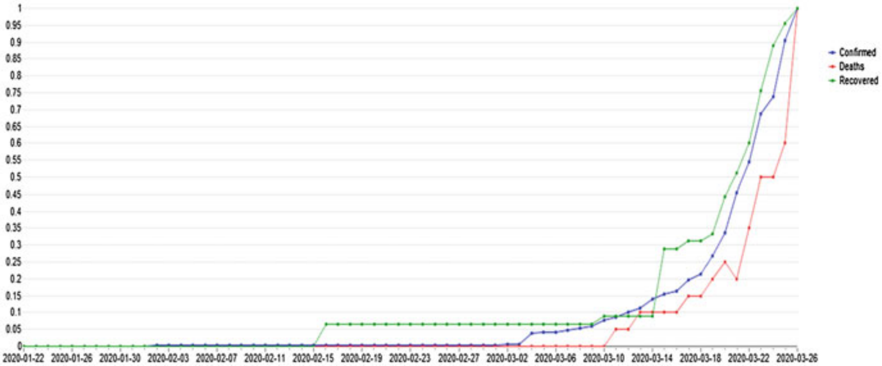


Fig. 5.1 Time series plot of INDIA COVID-19

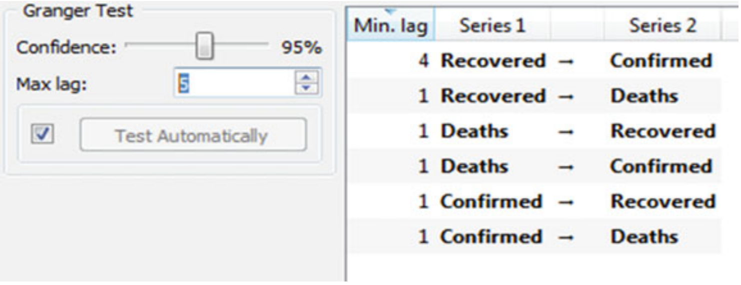
(Van Der Aalst 2016; Keller et al. 1994; Sui 2019; Mirkin 2019a). Table 5.1 provides the statistics of the considered dataset for the forecasting purpose. No missing value is the added advantage that helps in good prediction (Evans et al. 2007; Peat and Barton 2008).

Time series plot depicts the nature of the variable considered for the experimental purpose. It helps to recognize drifts in the data over spell. Intuitive in nature and provides faster insight about the changeover of the data over the time span. Line chart representation is used often that as higher clarity and informative in nature (Wen 2019). In the line chart representation parallelly visualize the changeover of the multiple variable of the dataset over the time at single shot.

Figure 5.1 shows the data connected for the variable like confirmed cases, death cases, and recovered cased of INDIA COVID-19 dataset on the normalize pattern. On applying the correlation over the dataset in orange data mining too we obtain the light of the correlation among the attributes. Higher correlation and minimum false

Table 5.2 Correlation of INDIA COVID-19 features

Data instances: 6				
Features: 2				Meta attributes: 2
	Feature 1	Feature 2	Correlation	FDR
1	Confirmed	Date	0.949	1.521012e − 30
2	Recovered	Date	0.937	4.52186e − 28
3	Confirmed	Recovered	0.847	1.43853e − 17
4	Deaths	Recovered	0.736	2.123428e − 11
5	Confirmed	Deaths	0.731	2.74335e − 11
6	Deaths	Date	0.694	5.7187e − 10



Granger Test

Confidence: 95%

Max lag: 5

☑ Test Automatically

Min. lag	Series 1		Series 2
4	Recovered	→	Confirmed
1	Recovered	→	Deaths
1	Deaths	→	Recovered
1	Deaths	→	Confirmed
1	Confirmed	→	Recovered
1	Confirmed	→	Deaths

Fig. 5.2 Granger test on INDIA COVID-19

discovery rate is the pattern that is visualized in this dataset (Mirkin 2019b). It's very obvious date plays great role in time series representation. Table 5.2 provides insight about correlation between two features along with measures.

Granger test is the statistical method for determining the influence of one time series with another time series feature considered for the experimental part. Famously it is called as granger causality since provided by academician Clive Granger. Find application in various fields like economics, neuro science and so on. With confidence 95%, for the dataset it's interpreted that confirmed cases four times ahead of recovered cases. Similarly as of now for the 2 months of time series dataset lag is not much inferred but as the length of data grows the lag may be skewed positively or negatively (Wen et al. 2019; Ghysels et al. 2016). Figure 5.2 shows the granger test with max lag as 5.

Curve fitting is the method of fabricating a curve based on the mathematical function. Obviously, that will make the perfect fit for the given data points with constraints. Fitted curve assist in the data visualization (Mudelsee 2019; Guest 2012; Maddams 1980). It helps in inferring values when data is missing with the generated function. Best fit can form in either straight manner or curve. Criterion value provides the insight about best fit. Figures 5.3, 5.4, and 5.5 illustrates the curve fit for the selected variable confirmed, death, and recovered, respectively.

Various measures help in understanding the accuracy of the fitted curve. Mathematical function for each variable is responsible for setting values in any part of

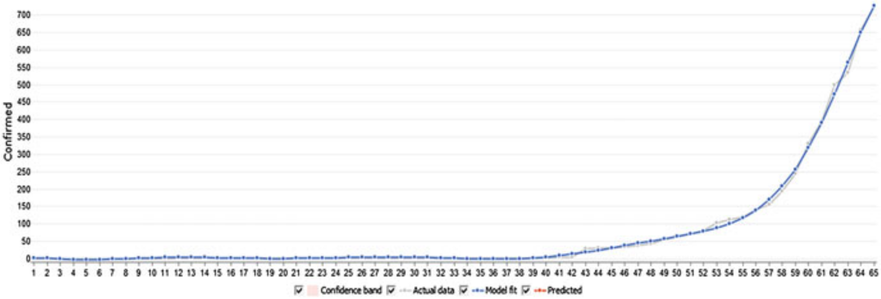


Fig. 5.3 Curve fitting of confirmed cases in INDIA COVID-19

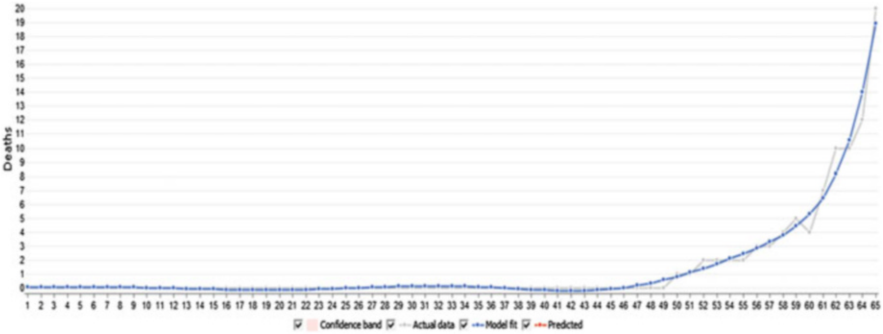


Fig. 5.4 Curve fitting of death cases in INDIA COVID-19

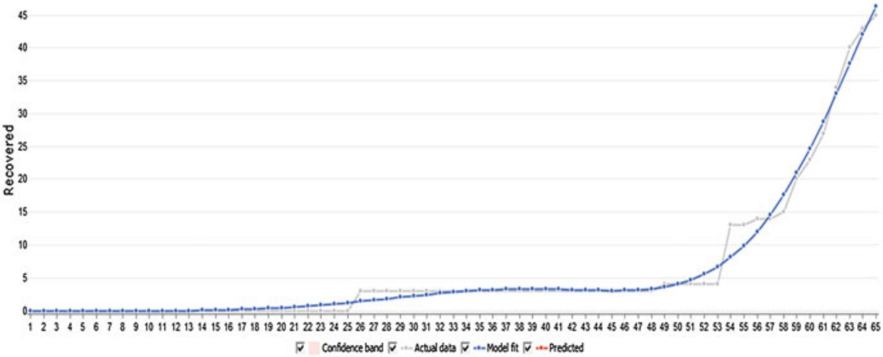
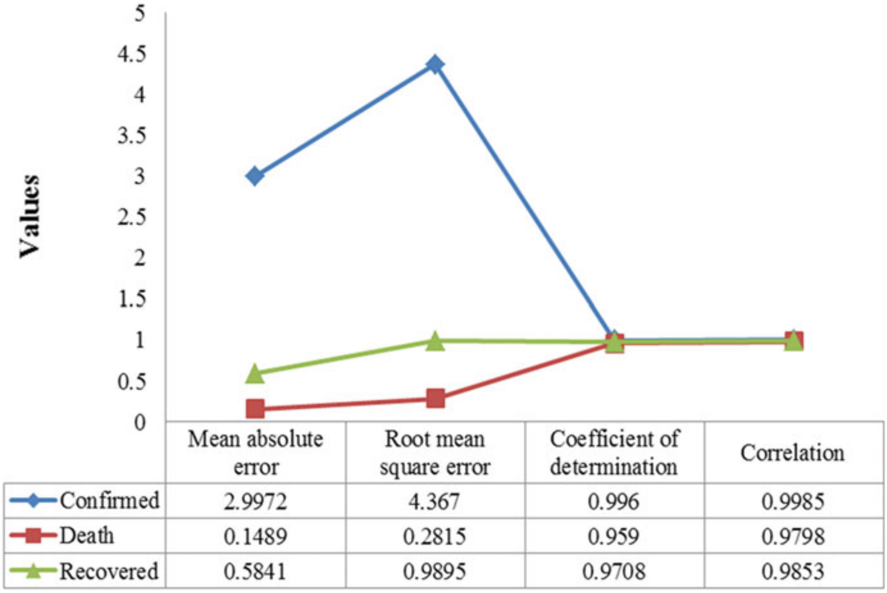


Fig. 5.5 Curve fitting of recovered cases in INDIA COVID-19

curve. Based on this perspective equations 5.3, 5.4, and 5.5 represents function of the confirmed, deaths, and recovered features, respectively.



	Mean absolute error	Root mean square error	Coefficient of determination	Correlation
◆ Confirmed	2.9972	4.367	0.996	0.9985
■ Death	0.1489	0.2815	0.959	0.9798
▲ Recovered	0.5841	0.9895	0.9708	0.9853

Fig. 5.6 Accuracy of model fit

$$\begin{aligned}Y_1 &= 2.51235 + \text{time}^{2*(-1.52766)} + \text{time}^{3*0.515268} + \text{time}^{4*(-0.0698097)} \\ &\quad + \text{time}^{5*0.00501064} + \text{time}^{6*(-0.000210204)} + \text{time}^{7*5.31865\text{e}-06} \\ &\quad + \text{time}^{8*(-7.98039\text{e}-08)} + \text{time}^{9*6.531\text{e}-10} + \text{time}^{10*(-2.243\text{e}-12)}\end{aligned} \tag{5.3}$$

$$\begin{aligned}Y_1 &= -0.0657009 + \text{time}^{5*3.21508\text{e}-07} + \text{time}^{7*(-3.58273\text{e}-10)} \\ &\quad + \text{time}^{9*1.77937\text{e}-13} + \text{time}^{10*(-1.67276\text{e}-15)}\end{aligned} \tag{5.4}$$

$$\begin{aligned}Y_1 &= 0.0795093 + \text{time}^{4*(-1.35655\text{e}-05)} + \text{time}^{6*1.22536\text{e}-07} \\ &\quad + \text{time}^{7*(-7.69865\text{e}-09)} + \text{time}^{8*1.98941\text{e}-10} + \text{time}^{9*(-2.38232\text{e}-12)} \\ &\quad + \text{time}^{10*1.09322\text{e}-14}\end{aligned} \tag{5.5}$$

Ranking of the best fit in the ascending order begins with deaths, recovered, and confirmed variable with criterion value as 0.0406, 0.0297 and 0.01181 for the mentioned functions respectively. Figure 5.6 graphical represents the plot of the mean absolute error (MAE), root mean square error (RMSE), coefficient of determination and correlation of model fit for the three variables of dataset.

We have conducted our experiment in GMDH based on four core algorithms namely Combinatorial (quick), Stepwise forward selection, Stepwise mixed selection, and GMDH neural network.

5.4.1 Combinatorial (Quick) Approach

The traditional combinatorial GMDH method produces models of all conceivable input variable mixes and chooses best model from the created set of models as indicated by a picked choice standard (Anastasakis and Mort 2001). Here for the experimental part with combinatorial (quick) method, the parameters used are reorder observation as Pseudo-random, validation strategy as *k*-fold validation, twofold, validation criteria as RMSE.balance, variable ranking as correlation, drop variable as rank 5, additional variable as *xi.xj* with return best model as 100 with time series mode.

$$Y_1(t) = -1497.69 + \text{``Confirmed}[t-8], \text{cubert''} * 413.182 \tag{5.6}$$

With the help of mathematical function (5.6) and system generated criterion value of 0.00153, the system is predicting the confirmed cases. Figure 5.7 shows the predicted values for confirmed cases.

$$\begin{aligned} Y_1[t] = 1.84925 &+ \text{``Deaths}[t-5], \text{cubert''} * (-2.82812) \\ &+ \text{``Deaths}[t-9], \text{cubert''} * (-2.50114) + \text{cycle} * 1.17415 \end{aligned} \tag{5.7}$$

With the above mathematical function (5.7) and system generated criterion value of 0.11036, the system is predicting the death cases. Figure 5.8 shows the predicted values for death cases.

$$Y_1[t] = -98.82 + \text{cycle} * 1.44 \tag{5.8}$$

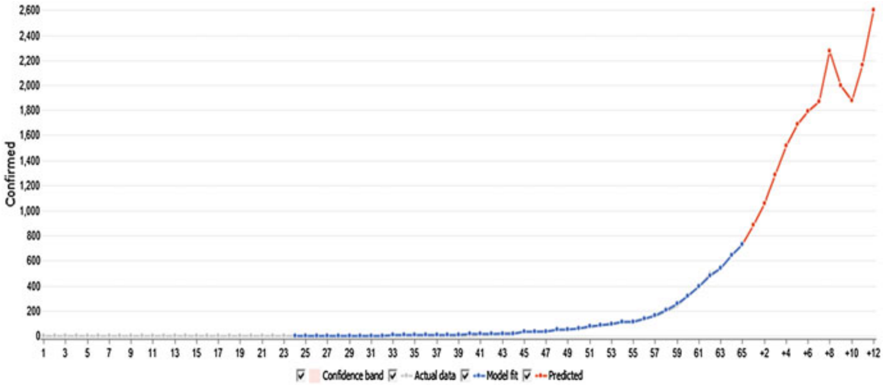


Fig. 5.7 Confirmed case prediction graph plot

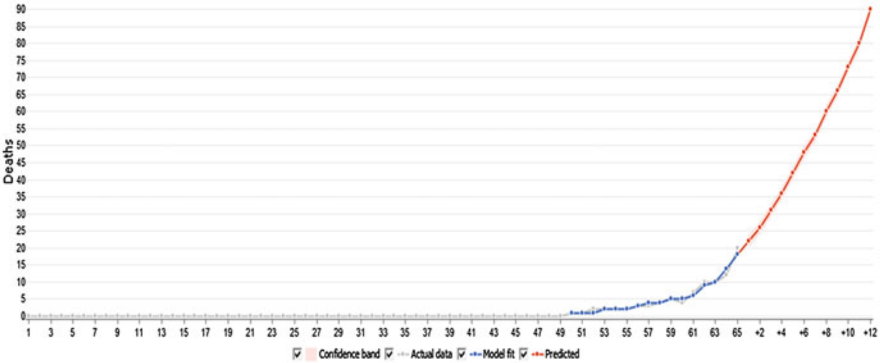


Fig. 5.8 Death case prediction graph plot

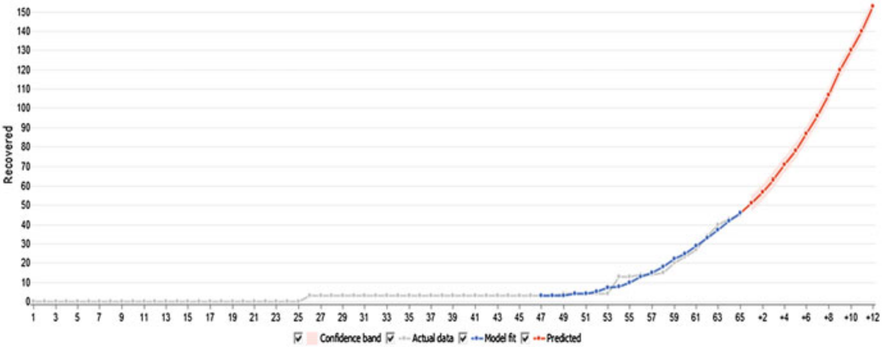


Fig. 5.9 Recovered case prediction graph plot

With the above mathematical function (5.8) and system generated criterion value of 0.17355, the system is predicting the recovered cases. Figure 5.9 shows the predicted values for recovered cases.

Tables 5.3 and 5.4 shows the predicted value and post-processed results of confirmed, death and recovered cases based on combinatorial (quick) algorithm.

5.4.2 Stepwise Forward Selection Approach

Forward selection is a kind of stepwise regression which starts with an unfilled model and includes variables individually. In each forward advance one can include the one variable that gives the absolute best improvement to your model (Glen 2019) . Here for the experimental part with stepwise forward selection method, the parameters used are reorder observation as Pseudo-random, validation strategy as *k*-fold validation, twofold, validation criteria as RMSE.balance, variable ranking as

Table 5.3 Forecast based on combinatorial(quick) approach

Target name	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12
Confirmed	887	1056	1282	1516	1689	1790	1874	2275	1996	1879	2163	2606
Deaths	22	26	31	36	42	48	53	60	66	73	80	90
Recovered	51	57	63	71	78	87	96	107	120	130	140	153

Table 5.4 Post-processed results by combinatorial (quick) algorithm

Post-processed results	Model fit (confirmed)	Model fit (death)	Model fit (recovered)
MAE	5.54	0.56	1.57
RMSE	6.89	0.90	2.05
Standard deviation of residuals (SD)	6.85	0.89	2.05
Correlation	0.99	0.98	0.98

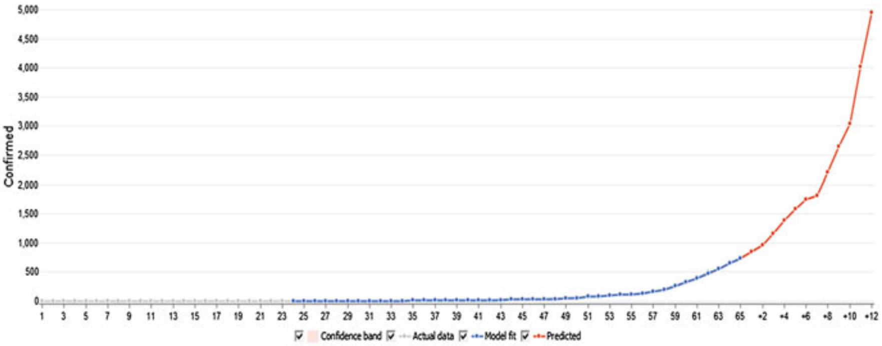


Fig. 5.10 Confirmed case graph plot

correlation, drop variable as rank 100, no additional variable are used with limit model complexity as 200, return best model as 100 with time series mode.

$$Y_1[t] = -1497.69 + \text{``Confirmed}[t-8], \text{cubert''} * 413.182 \tag{5.9}$$

With the above mathematical function (5.9) and system generated criterion value of 0.00153, the system is predicting the confirmed cases. Figure 5.10 shows the predicted values for confirmed cases.

$$\begin{aligned} Y_1[t] = 1.84925 + \text{cycle} * 1.17415 + \text{``Deaths}[t-5], \text{cubert''} \\ * (-2.82812) + \text{``Deaths}[t-9], \text{cubert''} * (-2.50114) \end{aligned} \tag{5.10}$$

With the above mathematical function (5.10) and system generated criterion value of 0.11036, the system is predicting the death cases. Figure 5.11 shows the predicted values for death cases.

$$Y_1[t] = -98.82 + \text{cycle} * 1.44 \tag{5.11}$$

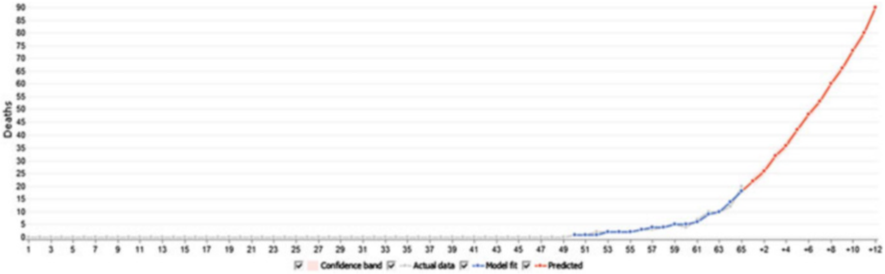


Fig. 5.11 Death case graph plot

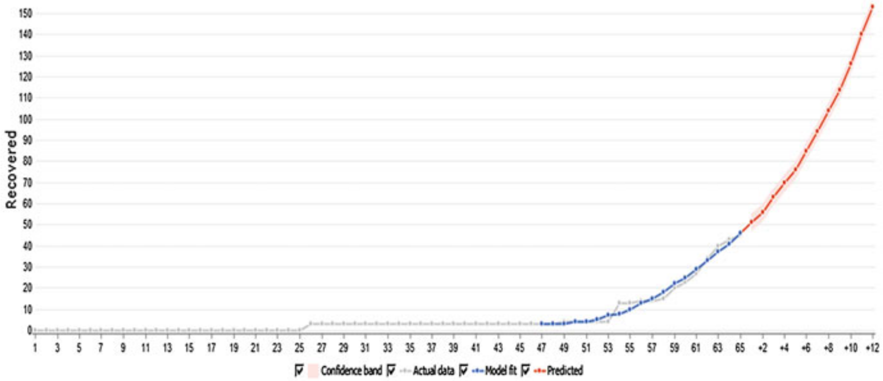


Fig. 5.12 Recovered case prediction graph plot

With the above mathematical function (5.11) and system generated criterion value of 0.17355, the system is predicting the recovered cases. Figure 5.12 shows the predicted values for recovered cases.

Tables 5.5 and 5.6 shows the predicted value and post-processed results of confirmed, death, and recovered cases based on stepwise forward selection algorithm.

5.4.3 Stepwise Mixed Selection Approach

The mixed stepwise variable determination system will ponder both including and evacuating one variable at each progression and make the best stride. In mixed calculation we could without much of a stretch include one variable, at that point include or expel another and afterward evacuate the primary variable included (CMU Statistics 2015). Here for the experimental part with stepwise mixed selection method, the parameters used are reorder observation as Pseudo-random, validation strategy as *k*-fold validation, twofold, validation criteria as RMSE.balance, variable

Table 5.5 Forecast based on stepwise forward selection approach

Target name	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12
Confirmed	856	972	1164	1392	1584	1750	1818	2212	2648	3040	4018	4952
Deaths	22	26	32	36	42	48	53	60	66	73	80	90
Recovered	51	56	63	70	76	85	94	104	114	126	140	153

Table 5.6 Post-processed results by stepwise forward selection algorithm

Post-processed results	Model fit (confirmed)	Model fit (death)	Model fit (recovered)
MAE	6.23	0.56	1.06
RMSE	8.26	0.90	1.47
Standard deviation of residuals (SD)	8.22	0.89	1.46
Correlation	0.99	0.98	0.99

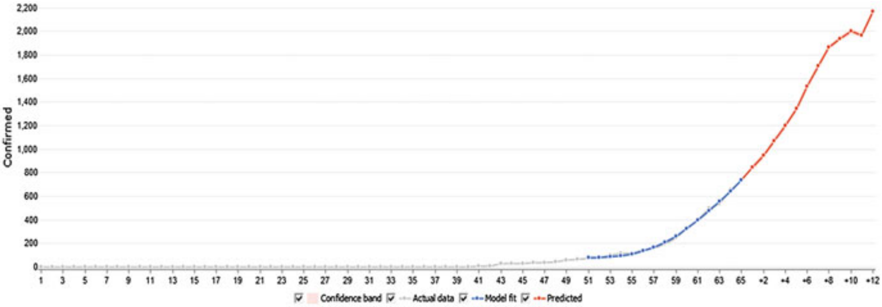


Fig. 5.13 Confirmed case prediction graph plot

ranking as correlation, drop variable as rank 100, no additional variable are used with limit model complexity as 200, return best model as 100 with time series mode.

$$Y_1[t] = 304.496 + N_3 * N_2 * 0.000803145 \tag{5.12}$$

$$N_2[t] = 301.867 + N_4 * N_3 * 0.000810118 \tag{5.13}$$

$$N_3[t] = 298.562 + N_5 * N_4 * 0.000818998 \tag{5.14}$$

$$N_4[t] = 296.225 + N_6 * N_5 * 0.000825283 \tag{5.15}$$

$$N_5[t] = 71.2228 + \text{time} * N_6 * 0.0161384 \tag{5.16}$$

$$N_6[t] = -1591.51 + \text{time} * \text{cycle} * 0.00773787 \tag{5.17}$$

With the above mathematical functions (5.12–5.17) and system generated criterion value of 0.0921, the system is predicting the confirmed cases. Figure 5.13 shows the predicted values for confirmed cases.

$$\begin{aligned} Y_1[t] = 1.84925 + \text{cycle} * 1.17415 + \text{``Deaths}[t-5], \text{cubert''} \\ * (-2.82812) + \text{``Deaths}[t-9], \text{cubert''} * (-2.50114) \end{aligned} \tag{5.18}$$

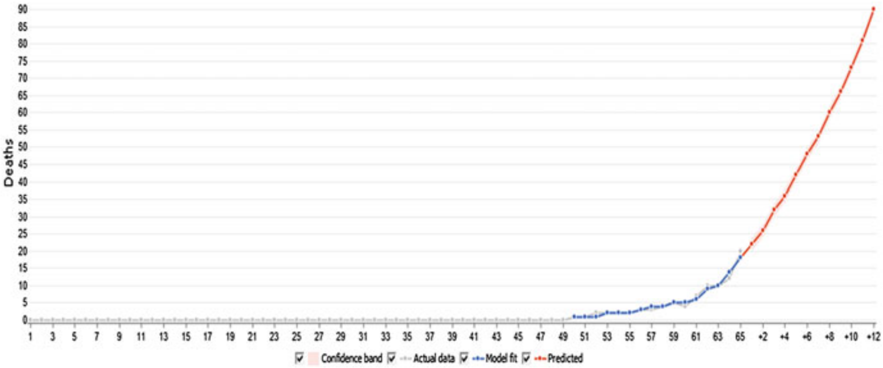


Fig. 5.14 Death case prediction graph plot

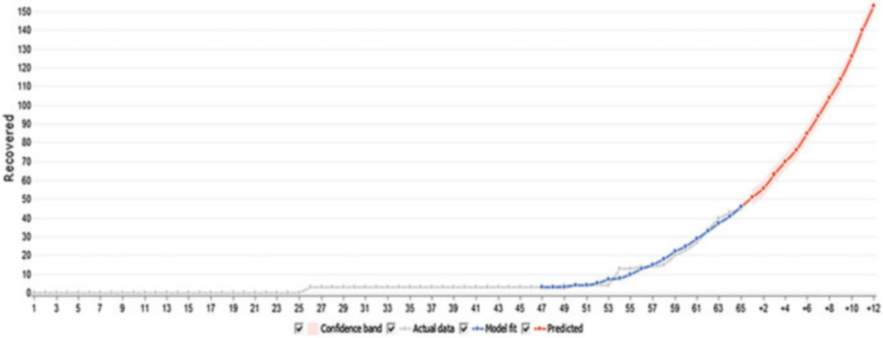


Fig. 5.15 Recovered case prediction graph plot

With the above mathematical function (5.18) and system generated criterion value of 0.11036, the system is predicting the death cases. Figure 5.14 shows the predicted values for death cases.

$$Y_1[t] = -98.82 + \text{cycle} * 1.44 \tag{5.19}$$

With the above mathematical function (5.19) and system generated criterion value of 0.17355, the system is predicting the recovered cases . Figure 5.15 shows the predicted values for recovered cases.

Tables 5.7 and 5.8 shows the predicted value and post-processed results of confirmed, death, and recovered cases based on stepwise mixed selection algorithm.

Table 5.7 Forecast based on stepwise mixed selection approach

Target name	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12
Confirmed	845	947	1071	1392	1199	1343	1706	1864	1939	2002	1968	2172
Deaths	22	26	32	36	42	48	53	60	66	73	81	90
Recovered	51	56	63	70	76	85	94	104	114	126	140	153

Table 5.8 Post-processed results by stepwise mixed selection algorithm

Post-processed results	Model fit (confirmed)	Model fit (death)	Model fit (recovered)
MAE	11.2	0.56	1.63
RMSE	13.06	0.90	2.09
Standard deviation of residuals (SD)	13.05	0.89	2.08
Correlation	0.99	0.98	0.98

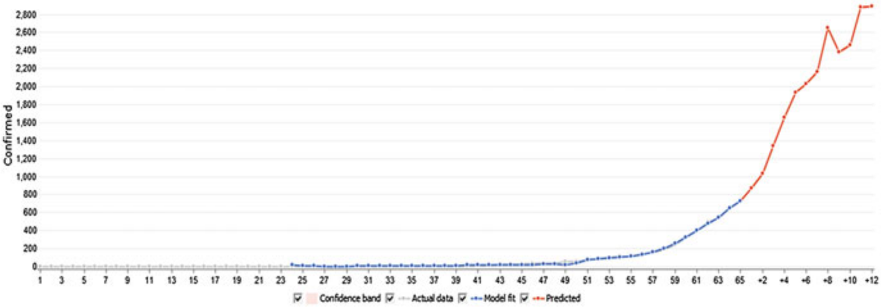


Fig. 5.16 Confirmed case prediction graph plot

5.4.4 GMDH Neural Network Approach

GMDH neural network comprehends time arrangement anticipating and information mining undertakings by building artificial neural networks and applying them to the information. Neural network estimating is more adaptable than ordinary linear or polynomial approximations and is along these lines progressively exact. With neural networks one can find and consider nonlinear associations and connections among information and construct an up-and-comer model with high forecast quality (NNS 2020) . Here for the experimental part with stepwise mixed selection method, the parameters used are reorder observation as Pseudo-random, validation strategy as *k*-fold validation, twofold, validation criteria as RMSE.balance, variable ranking as correlation, drop variable as rank 100, neuron function as $a + xi$ (linear), maximum number of layers as 33 with initial layer width as 1 with time series mode.

$$Y_1[t] = -1497.69 + {}^{``}\text{Confirmed}[t-8], \text{cubert}^{"} * 413.182 \tag{5.20}$$

With the above mathematical function (5.20) and system generated criterion value of 0.0015284, the system is predicting the confirmed cases. Figure 5.16 shows the predicted values for confirmed cases.

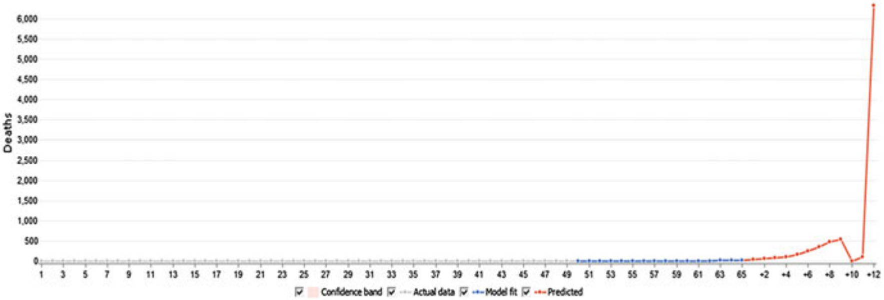


Fig. 5.17 Death case prediction graph plot

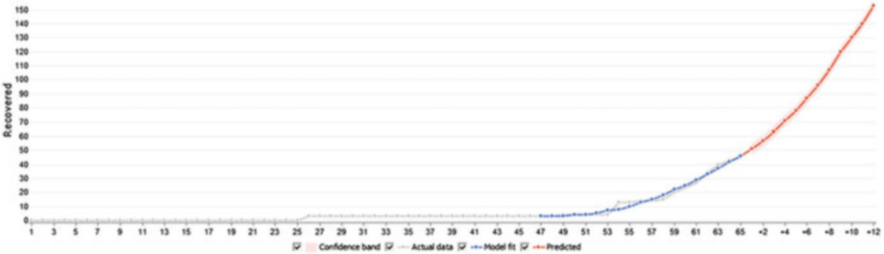


Fig. 5.18 Recovered case prediction graph plot

$$Y_1[t] = 0.426828 + ``\text{Deaths}[t-6].\text{cubert}" * N_3 * 0.142041 + N_3 * 0.740675 \tag{5.21}$$

$$N_3[t] = -0.219251 + ``\text{Deaths}[t-5].\text{cubert}" * N_3 * 0.491739 + N_3 * 0.957112 \tag{5.22}$$

$$N_4[t] = 0.438937 + ``\text{Deaths}[t-9].\text{cubert}" * N_5 * 0.126892 + N_5 * 0.797122 \tag{5.23}$$

$$N_5[t] = 0.671779 + ``\text{Deaths}[t-1].\text{cubert}" * \text{cycle} * 0.437579 \tag{5.24}$$

With the above mathematical functions (5.21–5.24) and system generated criterion value of 0.031344, the system is predicting the death cases. Figure 5.17 shows the predicted values for death cases.

$$Y_1[t] = -98.82 + \text{cycle} * 1.44 \tag{5.25}$$

With the above mathematical function (5.25) and system generated criterion value of 0.17355, the system is predicting the recovered cases. Figure 5.18 shows the predicted values for recovered cases.

Tables 5.9 and 5.10 shows the predicted value and post-processed results of confirmed, death, and recovered cases based on GMDH-NN algorithm.

5.5 Comparison Between the Algorithms Based on MAE, RMSE, SD, Correlation

Figure 5.19 shows the comparison of various used algorithms on parameters like Correlation, SD, MAE, RMSE. Based on the comparison it is clear that the stepwise mixed algorithm gives good prediction result for confirmed cases.

Figure 5.20 shows the comparison of various used algorithms on parameters like Correlation, SD, MAE, RMSE. Based on the comparison it is clear that the GMDH-NN algorithm gives good prediction result for death cases.

Figure 5.21 shows the comparison of various used algorithms on parameters like Correlation, SD, MAE, RMSE. Based on the comparison it is clear that the GMDH-NN algorithm gives good prediction result for death cases.

5.6 Conclusion

As this disease is declared as an epidemic, the present study will help researchers to understand the impact of this outbreak. We have used combinatorial (quick), stepwise forward selection, stepwise mixed selection and GMDH neural network to predict the spread of disease in India. Mathematical function mentioned in the each approach provides insight about the provided prediction. From the parametric comparisons it is clear that the GMDH-NN algorithm provides good accuracy in our case. Post-processed results obtained give the accuracy of the fitted model. COVID-19 provides a broad spectrum of future work in various disciplines.

Table 5.9 Forecast based on GMDH-NN approach

Target name	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12
Confirmed	869	1038	1337	1655	1933	2029	2161	2653	2380	2460	2877	2889
Deaths	22	27	32	36	43	48	53	60	66	73	81	90
Recovered	51	57	63	71	78	87	96	107	120	130	140	153

Table 5.10 Post-processed results by GMDH-NN approach

Post-processed results	Model fit	Model fit	Model fit
MAE	8.64	0.12	1.68
RMSE	10.56	0.35	2.20
Standard deviation of residuals (SD)	10.55	0.35	2.19
Correlation	0.99	0.99	0.98

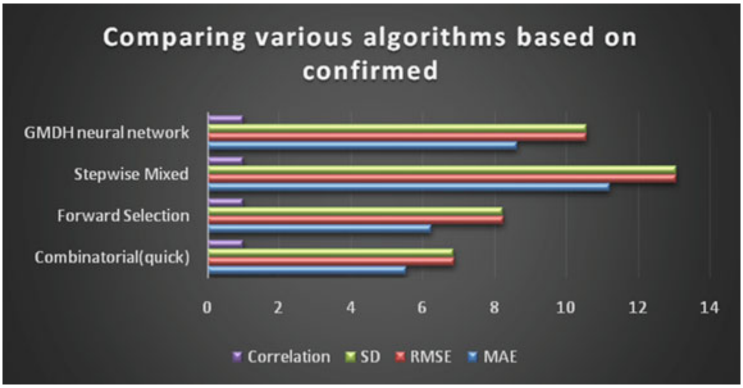


Fig. 5.19 Comparing algorithms for confirmed cases based on various parameters

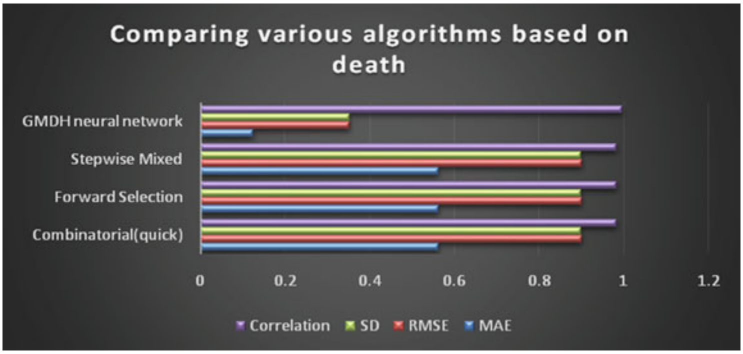


Fig. 5.20 Comparing algorithms for death cases based on various parameters

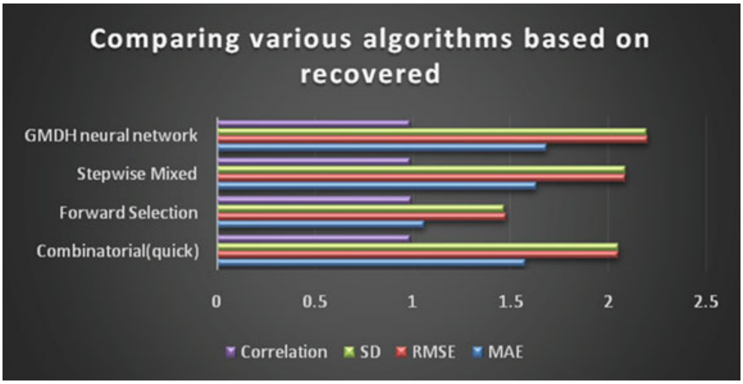


Fig. 5.21 Comparing algorithms for recovered cases based on various parameters

References

Abdolrahimi S, Nasernejad B, Pazuki G (2014) Prediction of partition coefficients of alkaloids in ionic liquids based aqueous biphasic systems using hybrid group method of data handling (GMDH) neural network. J Mol Liq 191:79–84

Ahmadi H, Mottaghitalab M, Nariman-Zadeh N (2007) Group method of data handling-type neural network prediction of broiler performance based on dietary metabolizable energy, methionine, and lysine. J Appl Poult Res 16(4):494–501

Anastasakis L, Mort N (2001) The development of self-organization techniques in modelling: a review of the group method of data handling (GMDH). Research Report-University of Sheffield Department of Automatic Control and Systems Engineering

Atashrouz S, Pazuki G, Kakhki SS (2015) A GMDH-type neural network for prediction of water activity in glycol and poly (ethylene glycol) solutions. J Mol Liq 202:95–100

Bioinformatics Laboratory (2020) Data mining. https://orange.biolab.si/. Accessed 19 Mar 2020

Chen H, Guo J, Wang C, Luo F, Yu X, Zhang W et al (2020) Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. Lancet 395(10226):809–815

CMU Statistics (2015). https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/26/lecture-26.pdf

Dorn M, Braga AL, Llanos CH, Coelho LS (2012) A GMDH polynomial neural network-based method to predict approximate three-dimensional structures of polypeptides. Expert Syst Appl 39(15):12268–12279

Evans JR, Olson DL, Olson DL (2007) Statistics, data analysis, and decision modeling. Pearson/Prentice Hall, Upper Saddle River, NJ

Ghanadzadeh H, Ganji M, Fallahi S (2012) Mathematical model of liquid–liquid equilibrium for a ternary system using the GMDH-type neural network and genetic algorithm. Appl Math Model 36(9):4096–4105

Ghysels E, Hill JB, Motegi K (2016) Testing for Granger causality with mixed frequency data. J Econ 192(1):207–230

Glen S (2019) Forward selection: definition from StatisticsHowTo.com. Elementary Statistics for the rest of us! https://www.statisticshowto.com/forward-selection/

GMDH (n.d.) Group Method of Data Handling (GMDH) for deep learning, data mining algorithms optimization, fuzzy models' analysis, forecasting neural networks, and modelling software systems. https://www.gmdh.net/

Guest PG (2012) Numerical methods of curve fitting. Cambridge University Press, Cambridge

Ivakhnenko AG (1971) Polynomial theory of complex systems. IEEE Trans Syst Man Cybern 4:364–378

Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S et al (2020) COVID-19 patient health prediction using boosted random forest algorithm. Front Public Health 8:357

Keller PR, Keller MM, Markel S, Mallinckrodt AJ, McKay S (1994) Visual cues: practical data visualization. Comput Phys 8(3):297–298

KP (2020) COVID-19 dataset. Kaggle: your machine learning and data science community. https://www.kaggle.com/imdevskp/corona-virus-report/data

Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J (2020) The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med 27(2):taaa021

Maddams WF (1980) The scope and limitations of curve fitting. Appl Spectrosc 34(3):245–267

Mahtani S (2020) WHO director-general's opening remarks at the media briefing on COVID-19—11 March 2020. The Washington Post. https://www.washingtonpost.com/world/asia_pacific/coronavirus-china-live-updates/2020/02/05/114ced8a-479c-11ea-bc78-8a18f7afcee7_story.html. Accessed 5 Feb 2020

Mirkin B (2019a) Core data analysis: summarization, correlation, and visualization. Springer, Cham

Mirkin B (2019b) Learning correlations. In: Core data analysis: summarization, correlation, and visualization. Springer, Cham, pp 163–292

Mudelsee M (2019) Trend analysis of climate time series: a review of methods. Earth Sci Rev 190:310–322

Naderpour H, Nagai K, Fakharian P, Haji M (2019) Innovative models for prediction of compressive strength of FRP-confined circular reinforced concrete columns using soft computing methods. Compos Struct 215:69–84

Najafzadeh M (2015) Neuro-fuzzy GMDH based particle swarm optimization for prediction of scour depth at downstream of grade control structures. Eng Sci Technol 18(1):42–51

Nguyen TN, Lee S, Nguyen-Xuan H, Lee J (2019) A novel analysis-prediction approach for geometrically nonlinear problems using group method of data handling. Comput Methods Appl Mech Eng 354:506–526

NNS (2020) Best neural network software (NNS) in 2020 (Free academic license). GMDH. https://gmdhsoftware.com/neural-network-software. Accessed 19 Mar 2020

Novel CPERE (2020) The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi 41(2):145

Onwubolu GC (ed) (2009) Hybrid self-organizing modeling systems, vol 211. Springer, Berlin

Pan F, Ye T, Sun P, Gui S, Liang B, Li L et al (2020) Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia. Radiology 200370

Pazuki G, Kakhki SS (2013) A hybrid GMDH neural network to investigate partition coefficients of Penicillin G Acylase in polymer–salt aqueous two-phase systems. J Mol Liq 188:131–135

Peat J, Barton B (2008) Medical statistics: a guide to data analysis and critical appraisal. Wiley, New York

Sui Z (2019) Social media text data visualization modeling: a timely topic score technique. Am J Manage Sci Eng 4(3):49–55

Sujatha R, Chatterjee JM, Hassanien AE (2020a) A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 34:959–972. https://doi.org/10.1007/s00477-020-01827-8

Sujatha R, Chatterjee J, Hassanien AE (2020b) A machine learning methodology for forecasting of the COVID-19 cases in India

Van Der Aalst W (2016) Data science in action. In: Process mining. Springer, Berlin, pp 3–23

Wen Q (2019) Asset growth and stock market returns: a time-series analysis. Rev Finance 23(3):599–628

Wen X, Bai Z, Li Y, Liu S, Wang F (2019) Correlation analysis of competitive electricity market based on granger test

WHO (2020) WHO director-general's opening remarks at the media briefing on COVID-19—11 March 2020. WHO | World Health Organization. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19%2D%2D-11-march-2020

World Health Organization (2020a) Statement on the second meeting of the international health regulations emergency committee regarding the outbreak of novel coronavirus (2019-nCoV). WHO | World Health Organization. https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov). Accessed 30 Jan 2020

World Health Organization (2020b) Coronavirus disease 2019 (COVID-19): situation report, 47. World Health Organization. https://apps.who.int/iris/handle/10665/331444

Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C et al (2020) Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respir Med 8(4):420–422