The Pharma Innovation Journal 2023; 12(2): 3316-3320

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2023; 12(2): 3316-3320
© 2023 TPI
www.thepharmajournal.com
Received: 09-12-2022
Accepted: 12-01-2023

Ambreen Hamadani
National Institute of Technology, Srinagar, Jammu and Kashmir, India

Onyijen Ojei Harrison
Glorious Vision University, Edo, Nigeria

Nazir A Ganai
Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir, Jammu and Kashmir, India

Israel Ehizuelen Ebhohimen
Ambrose Alli University, Ekpoma, Nigeria

Oluwatobi Samuel Awe
Glorious Vision University, Edo, Nigeria

Celestine Uche Agwi
Glorious Vision University, Edo, Nigeria

J Bashir
National Institute of Technology, Srinagar, Jammu and Kashmir, India

Corresponding Author:
Ambreen Hamadani
National Institute of Technology, Srinagar, Jammu and Kashmir, India

Exploration of machine learning algorithms for the evaluation of factors affecting COVID-19 death rates

Ambreen Hamadani, Onyijen Ojei Harrison, Nazir A Ganai, Israel Ehizuelen Ebhohimen, Oluwatobi Samuel Awe, Celestine Uche Agwi and J Bashir

Abstract
The COVID-19 pandemic had a great impact on the world. It wreaked a great havoc around the globe impacting every human on the face of this planet in one way or the other. The unprecedented challenges that the world faced demonstrated the need for big data and its mining for making futuristic predictions so that disasters such as this could be averted, especially when human lives are at stake. The present study was therefore undertaken to harness the potential of artificial intelligence algorithms for drawing useful inferences from the COVID 19 dataset by "Our World in Data". Appropriate data preparation was done using data imputation, encoding, and normalization. Feature selection was performed on the 67 initial features. Ordinary least squares and artificial neural networks (ANN) were used in this research. Convolutional Neural Networks (CNN) were also compared with the feed-forward back propagation algorithm for the current dataset. Our results indicate that the total cases, total tests, total vaccinations, and diabetes prevalence had the highest feature scores of 10983.38, 6636.90, 5118.28, and 2150.80 respectively. The coefficient of determination value for the regression equation was 0.972 indicating a good fit. For the ANN model with only one feature i.e., total vaccination, the correlation coefficient was 0.865 going on to show that vaccination was an important factor in preventing deaths. Our results also indicate that both the ANN and CNN had high prediction correlations (0.99), but the CNN had a lower error rate for the same. Additionally, the high feature scores of factors like total cases and total vaccinations influence the death rates greatly. Therefore, preventive measures like social distancing, masking up and vaccination could potentially reduce the death rate associated with COVID-19. We conclude that the machine learning models developed were accurate based on the correlation coefficient and could be used for drawing useful patterns regarding deaths during this great pandemic of the Century.

Keywords: Artificial intelligence, corona virus, COVID 19, mortality, vaccination, co-morbidity

Introduction
The COVID 19 pandemic caused mayhem as the world has hardly ever seen before. The unprecedented challenges that the world faced demonstrated the need for big data and its mining for making futuristic predictions so that disasters such as this could be averted. Artificial intelligence through big data mining is potentially transforming the world as we see it today, and it is being used in all spheres of life [1] including the fight against the pandemic ever since the beginning of 2020. Therefore, this area of advanced statistics could potentially help in the control as well as the prevention of COVID-19 [2]. This would also be useful for predicting other pandemics that may occur in the future. It could also prevent such pandemics from occurring in the future.

Data that has been generated ever since the COVID-19 pandemic is a goldmine, as the potential of AI can only be harnessed using good and reliable datasets. It is this data that is helping in understanding the various factors that predispose the population to death due to COVID-19 and artificial intelligence has been used by various authors for predicting various aspects of the COVID-19 pandemic [3-6].

Ever since the pandemic, the focus of scientists around the globe has been to save lives and reduce the morbidity and subsequently the mortality as much as possible. This research was therefore undertaken to understand the various factors influencing death rates due to COVID 19 using the COVID 19 dataset by "Our World in Data". An attempt to understand factors complicating the disease and leading to mortality has also been made.

Materials and Methods

Dataset

An open-source dataset for COVID-19 kept in public domain by "Our World in Data" was used in this study [7]. The original data contained 58,46,090 data points. The original features in the dataset included features like ISO code, location, continent, date, new cases, new cases smoothed, total deaths, new deaths, total cases, total cases/million, new cases/million, new cases smoothed/million, total deaths/ million, new deaths/million, new deaths smoothed/ million, new deaths smoothed, reproduction rate, ICU patients, ICU patients/million, hospital patients, hospital patients/million, weekly ICU admissions, weekly ICU admissions/million, weekly hospital admissions, weekly hospital admissions/million, new tests, tests, tests/thousand, new tests/ thousand, new tests smoothed, new tests smoothed/thousand, positive rate, tests/case, tests units, total vaccinations, people vaccinated, people fully vaccinated, total boosters, vaccinations, vaccinations smoothed, total vaccinations/ hundred, vaccinated/hundred, fully vaccinated/hundred, total boosters/hundred, new vaccinations smoothed/million, new vaccinated persons smoothed, new people vaccinated smoothed/hundred, stringency index, population density, median age, age: 65 older, age: 70 older, GDP/capita, extreme poverty, cardiovascular death rate, diabetes prevalence, female and male smokers, hand washing facilities, hospital beds/thousand, human development index, life expectancy, excess mortality cumulative absolute, excess mortality, excess mortality cumulative, excess mortality cumulative/million.

Data Preparation

For data preparation, the rows with missing variables were removed. All data was appropriately encoded and normalized. Normalization for all the datasets used in this study was done in the Python language [8] to achieve proper data transformation with the distribution having a mean of 0 as well as a standard deviation of 1.

Feature Selection Scores

Dimensionality reduction on the dataset was done by reducing the feature numbers and thereby the number of planes in the dataset, feature selection was done. Feature scores were derived based on an F-test estimate of the degree of linear dependency between two numerical variables: the input and the output. This was treated as a regression predictive modeling problem [8].

Ordinary Least Squares

A prediction equation for predicting total deaths was derived using ordinary least squares in Python. The models used had the following structure:

$$Y_{ik} = a + b_iX_i + b_hX_h + \ldots + b_k X_k + e_{ik},$$

Where Y_{ik} equals total deaths, a equals intercept, b_k equals regression coefficient estimated, X_k equal weight components and e_{ik} equals random error (NID (0, σ^2e)). The accuracy of this particular model was estimated by the coefficient of determination (R^2).

Neural Networks

Total deaths were used as labels for the neural networks. The number of features was kept variable while training the network so as to achieve optimization. The feature of total vaccinations was also used alone in the neural network to understand its impact on total deaths. The number of neurons/ layer, number of hidden layers, activation function, the learning rate, and optimizer settings were all heuristically determined. Convolutional Neural Networks were also compared with the feed forward back propagation algorithm. For this, the hyper parameters were also heuristically determined.

Results and Discussion

Feature Selection Scores

The feature selection scores of the features used in the dataset are given in table 1. Our results indicate that the total cases, total tests, total vaccinations, and diabetes prevalence had the highest feature scores of 10983.38, 6636.90, 5118.28, and 2150.80 respectively. The high feature selection scores of the diabetes prevalence indicate that people with diabetes have higher chances of causing mortality due to various comorbidities associated with the COVID-19 infection. This has also been widely reported in the literature [10, 11].

Table 1: Feature Selection scores of top features

Feature	Score
Total cases	10983.38
Total tests	6636.90
Total vaccinations	5118.28
Prevalence of Diabetes	2150.80
Hospital beds/thousand	215.53
GDP/capita	183.54
Hand washing facilities	151.23
Median age	112.77
Human development index	102.66
Male smokers	75.33
Population density	62.18
> 65 age	52.09
Life expectancy	39.40
>70 age	22.66
Heart related death rate	21.46
Female smokers	7.70
Reproduction rate	3.81
Extreme poverty	3.41
Stringency index	0.24

Ordinary Least Squares

The R2 value for the regression equation was 0.972 indicating a good fit. All features used in the equation were highly significant except for male smokers. This indicates that all factors have some role in the prediction of COVID-19 deaths. Total tests, age > 65 years, the heart related death rate of the country, hospital beds/thousand, life expectancy, and the human development index were all negatively correlated. The equation of all significant variables is given as

Y = 7.00e4 + 1.59e5x1 + 1059.20x2 - 9.41e4x3 + 2.94e4x4 + 3384.13x5 - 9246.62x6 + 6.09e4x7 - 2.12e5x8 + 1.24e5x9 + 1.88e4x10 + 7017.57x11 -2.17e4x12 + 2.61e4x13 + 1.24e4x14 +445.48x15 + 2.16e4x16 - 1.75e4x17 -4901.47x18 -8358.55x19.

Where x1 = total cases, x2 = reproduction rate, x3 = total tests, x4 = total vaccinations, x5 = the stringency index, x6 = density, x7 = age (median), x8 = >65 years age x9 = >70 years age,x10 = gdp/capita, x11 = extreme poverty, x12 = cardiovascular death rate, x13 = prevalence of diabetes, x14 = no. of female smokers, x15 = no. of male smokers, x16 =

availability of hand washing facilities, x17 = hospital beds/thousand, x18 = expectancy of life, x19 = human development index.

Another researcher [12] used various regression techniques to investigate the effects of various factors of health and their statistical and spacial association with COVID-19 and found that an R2-score of 0.60 for the regression equation is much lower than the one obtained by our study. A much higher R2 value (0.99) was reported [13] for the prediction of deaths in Nigeria due to COVID-19. The effect of age was found to be significant in the present study. It has also been reported most deaths have occurred in a particular age group i.e. 60-69 years [14].

Linear regression has also been used to predict death numbers in India [15] and the 5 week average death count has been predicted to be 211 with a 95% CI. Regression methodology was also used in Ethiopia to predict deaths with reasonable accuracy [16]. Multivariate linear regression has been reported to give an R2 score of 0.992 [21] which agrees with the findings of this study.

Neural Networks

The results obtained by the artificial neural networks and convolutional neural networks are given in table 2. For the ANN model with only one feature i.e., total vaccination, the correlation coefficient was 0.865 going on to show that deaths could be predicted using vaccination data alone with reasonable accuracy (Figure 2). It has also been reported in the literature that vaccination is most effective for the prevention of severe cases and deaths from COVID-19. A 75.31% and a 74.89% decrease in COVID cases and death rates respectively have also been reported among fully vaccinated people [17] which agrees with our study.

The learning rate for the ANN was optimized at 0.01. Our results indicate that both the ANN and CNN could predict the total deaths with a high correlation, but the CNN had a lower error rate for the same. High correlation was also obtained by [23, 24] for making predictions. The hyper parameters and the values obtained by the two models are given in table 2. Figure 3 gives the plot of the predicted versus test labels for the convolutional neural networks.

Table 2: Hyper parameter optimization for prediction of breeding values

	ANN	CNN
Num hidden layers	3	4 + 1 Conv1D + 1flatten
Neurons/layer	900	200
Batch size	100	50
Activation	swish	swish
Optimizer	ADAM	ADAM
Iterations	500	1000
Test MAE	6046.83	5071.24
Correlation	0.999	.999
Epoch	28	108

In consonance with our findings, it has been also found neural network models employing advanced statistics are useful for the forecast of the COVID related mortality rate [18] (Dhamodharavadhani, *et al.* 2020). However, random forest models have been reported to provide a lower accuracy of 94% for the predictions [19] than that reported in the present study. Artificial neural networks for modeling novel coronavirus incidence rates in the United States have shown that the error values were minimal while using this technique [20]. The performance of CNNs for COVID-19 disease detection has also been evaluated [22] and found its accuracy as high as 99.5%.

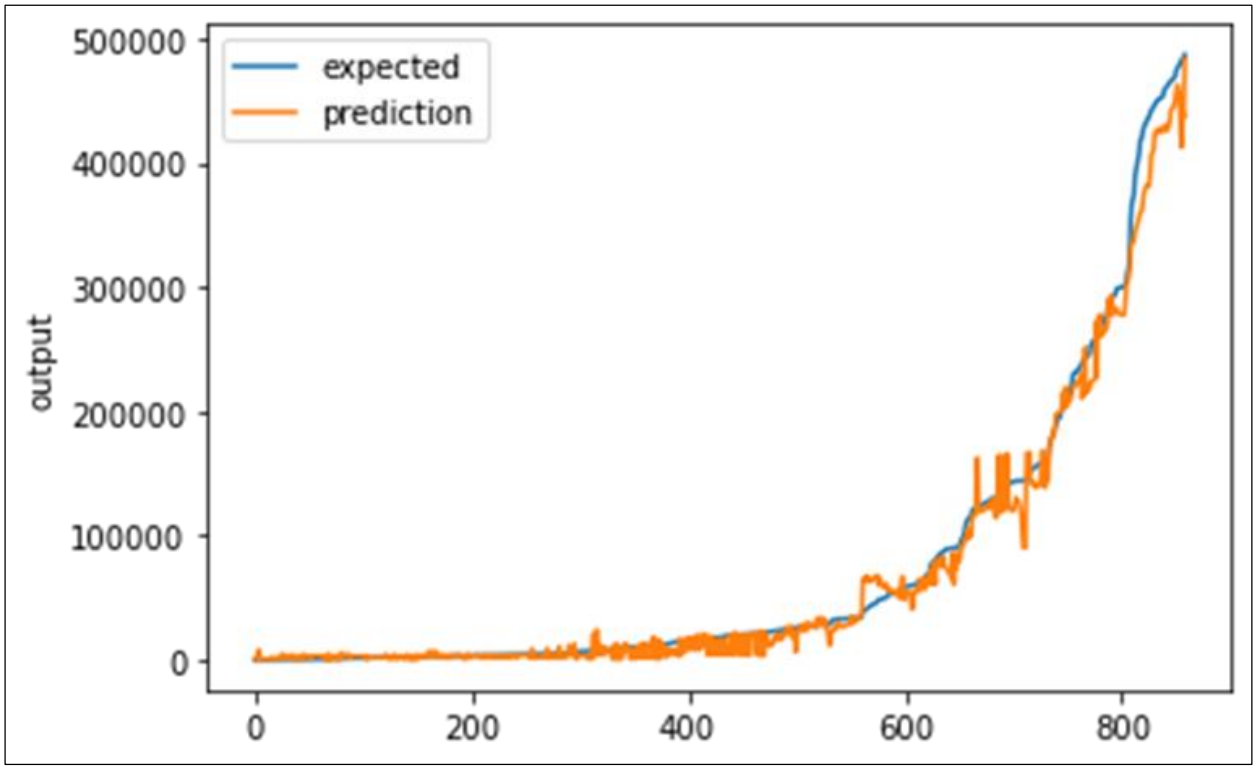

Fig 1: chart of predicted v/s test labels for ANN

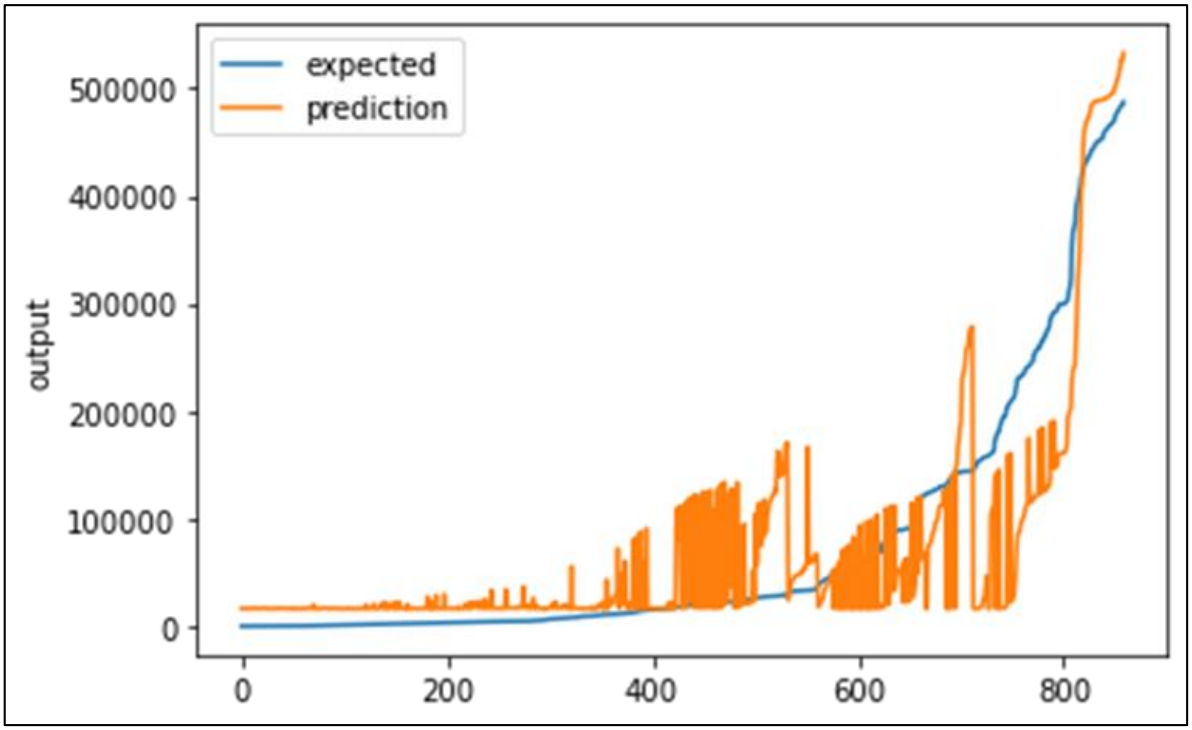

Fig 2: chart of predicted v/s test labels for ANN with only total vaccinations as a training feature

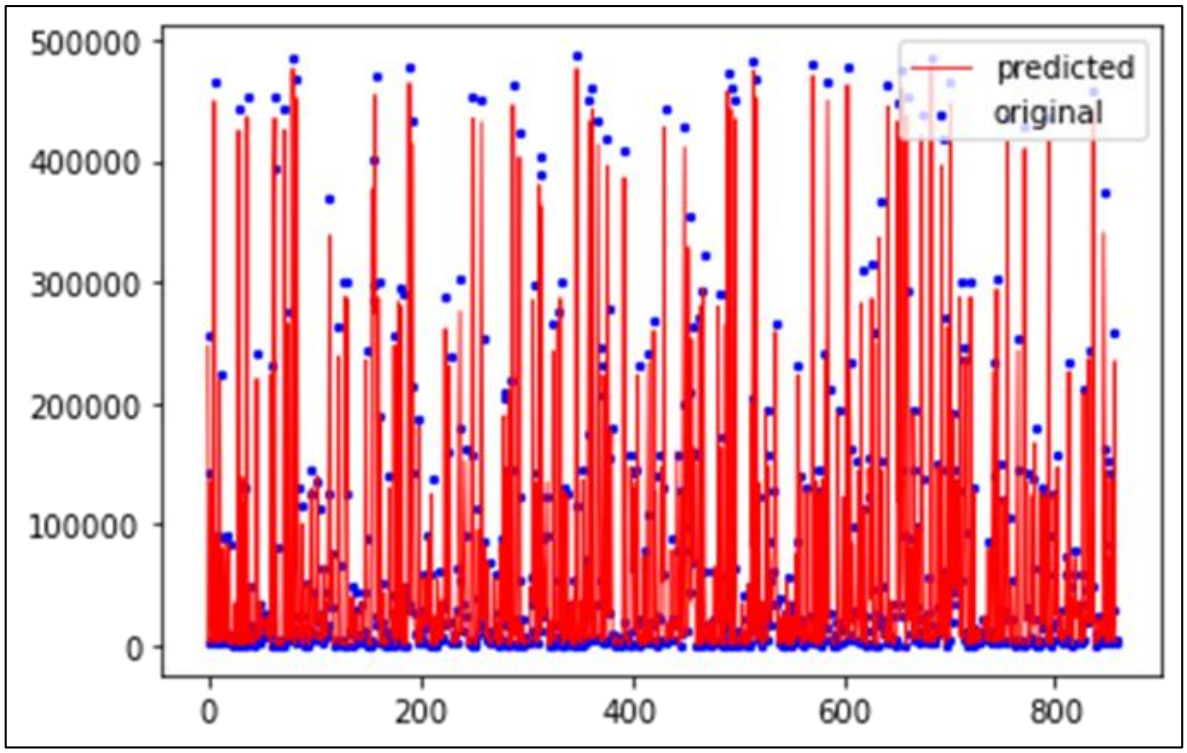

Fig 3: Plot of predicted versus test labels for the convolutional neural networks.

Conclusion

It is concluded that machine learning models developed were accurate based on the correlation coefficient and could be used for drawing useful patterns regarding deaths during the COVID-19 pandemic. Additionally, the high feature scores of factors like total cases and total vaccinations influence the death rates greatly. Therefore, the data indicate that preventive measures like social distancing, masking up and vaccination have in the past reduced the rate of spread and subsequently the deaths associated with COVID-19.

References

1. Hamadani A, Ganai NA, Andrabi SM, Shanaz S, Alam S. Evaluation of Artificial Intelligence Algorithms for the Prediction of Genetic Merit. Preprint; c2021 https://doi.org/10.21203/rs.3.rs-1488946/v1
2. Shamout FE, Shen Y, Wu N, Kaku A, Park J, Makino T, *et al*. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. NPJ Digit. Med, 2021;4:80. https://doi.org/10.1038/s41746-021-00453-0

3. Kundu S, Elhalawani H, Gichoya JW, Kahn Jr CE. How might ai and chest imaging help unravel COVID-19's mysteries? Radiol. Artif. Intell. 2021;2:e200053.
4. Khan AI, Shah JL, Bhat MM. Coro Net: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images. Comput. Meth. Prog. Bio. 2020;196:105581.
5. Ucar F, Korkmaz D. COVID Diagnosis-net: deep Bayes-squeeze Net based diagnostic of the coronavirus disease 2019 (COVID-19) from X-ray images. Med. Hypothese. 2019;140:109761.
6. Singh A, Gupta R, Ghosh A, Misra A. Diabetes in COVID-19: Prevalence, pathophysiology, prognosis, and practical considerations. Diabetes and Metabolic Syndrome: Clinical Research and Reviews. 2020;14(4):303-310. https://doi.org/10.1016/j.dsx.2020.04.004.
7. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, *et al*. Coronavirus Pandemic (COVID-19). Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]; c2020.
8. Pedregosa Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine Learning in {P}ython. Journal of Machine Learning Research. 2021;12:2825-2830
9. Singh D, Kumar V, Kaur M. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks. Eur. J. Clin. Microbiol. 2020;39:1379-1389.
10. Yang JK, Feng Y, Yuan MY, Yuan SY, Fu HJ, Wu BY, *et al*. Plasma glucose levels and diabetes are independent predictors for mortality and morbidity in patients with SARS. Diabet Med. 2006;23(6):623-628
11. Almalki A, Gokaraju B, Acquaah Y, Turlapaty A. Regression Analysis for COVID-19 Infections and Deaths Based on Food Access and Health Issues. Healthcare. 2022;10:324. https://doi.org/10.3390/ healthcare10020324
12. Onyijen OH, Hamadani A, Awojide S, Ebhohimen IE. Prediction of deaths from COVID-19 in Nigeria using various machine learning algorithms. Sau science-tech Journal. 2021;6(1):109-117.
13. Taib N, Baha RD, Teo AKJ, Kamarulzaman A, William T, Singh A, Mokhtar SA, *et al*. Characterisation of COVID-19 deaths by vaccination types and status in Malaysia between February and September 2021. The Lancet Regional Health–Western Pacific. 2021;18(100354):2666-6065
14. Ghosal S, Sengupta S, Majumder M, Sinha B. Linear Regression Analysis to predict the number of deaths in India due to SARS - CoV-2 at 6 weeks from day 0 to 100 cases. Diabetes and Metabolic Syndrome: Clinical Research and Reviews. 2021;14(4):311-315.
15. Argawu A. Regression Models for Predictions of COVID-19 New Cases and New Deaths in Ethiopia. International Journal of Theoretical and Applied Mathematics. 2021;6:53-63. 10.11648/j.ijtam.20200605.11.
16. Vanshika R, Monika B, Priya T, Rajiv A, Mohamed FA, Afzal H, *et al*. Analyzing the Effect of Vaccination over COVID Cases and Deaths in Asian Countries Using; Machine Learning Models. Frontiers in Cellular and Infection Microbiology; c2022. p. 11.
17. Dhamodharavadhani S, Rathipriya R, Chatterjee JM. COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models. Frontiers in Public Health; c2020. p. 8. https://www.frontiersin.org/article/10.3389/fpubh.2020.00441
18. Iwendi C, Bashir AK, Pasupuleti NS, Sujatha R, Chatterjee JM, Peshkar A. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Frontiers in Public Health. 2020;8:357.
19. Mollalo A, Mao L, Rashidi P, Glass GE. A GIS-Based Artificial Neural Network Model for Spatial Distribution of Tuberculosis across the Continental United States. International Journal of Environment Resources. 2019;16:157.
20. Suganya R, Arunadevi R, Buhari SM. COVID-19 Forecasting using Multivariate Linear Regression. PREPRINT (Version 1) available at Research; c2021. https://doi.org/10.21203/rs.3.rs-71963/v1
21. Reshi AA, Rustam F, Mehmood A, Alhossan A, Alrabiah Z, Ahmad A, *et al*. An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification. Complexity; c2021. p. 6621607. https://doi.org/10.1155/2021/6621607
22. Hamadani A, Ganai NA. Development of a multi-use decision support system for scientific management and breeding of sheep, Scientific Reports. 2022;12(1):19. DOI: 10.1038/s41598-022- 379 24091-y
23. Hamadani A, Ganai NA, Mudasir S, Shanaz S, Alam S, Hussan I. Comparison of artificial intelligence algorithms and their ranking for the prediction of genetic merit in sheep. Sci. Rep. 2022;12:18726. https://doi.org/10.1038/s41598-022-23499-w