# COVID-19 DOES NOT CORRELATE WITH THE TEMPERATURE

J. Adrian[1], Y. Mingxuan[2,3], T. Ye[4] , P. Norouzzadeh[1], M.H. Amini[2], M.A. Salari[1], E. Snir[2], B. Rahmani[1]

[1]St. Louis University, Computer Science Department, St. Louis, MO, USA
[2]Washington University, Olin Business School, St. Louis, MO, USA
[3]University of Macau, Taipa, China

## *ABSTRACT*

*The coronavirus disease 2019 (Covid-19) is now a global health crisis. According to the World Health Organization Situation Report, the number of Covid-19 confirmed cases reached 10,185,374 globally as of June 30, 2020.With global temperatures rising, many people expect the warmer weather to eradicate the virus, much like with the severe acute respiratory syndrome (SARS) in 2003. But will this expectation come true? Will Covid-19 cases surge in the Fall? In this project, we evaluate the relationship between the number of Covid-19 confirmed cases and the temperature during the early part of the epidemic, from March 2020 to June 2020. The analysis is based on monthly data regarding Covid-19 cases and temperature from188 countries/regions. The correlation coefficient suggests that there is virtually no relationship. To evaluate this further, we use K-Means clustering and Random Forest Regression (RFR). K-means clusteringis used to divide the countries into 10 groups of confirmed cases. However, the groups differ substantially in monthly temperature. The RFR also found no correlation between temperature and Covid-19. We conclude that the Covid-19 has no relationship with temperature.*

## *KEYWORDS*

*Covide-19, Temperature, Correlation, K-Means Clustering*

## 1. INTRODUCTION

The coronavirus disease 2019 (Covid -19) first appeared in Wuhan, China. As of March 3, 2020, 88.37% of the 90,870 worldwide reported cases were in China [1],[2],[3]. The WHO (World Health Organization) declared Covid-19 a global pandemic. By July 27, 2020, there were more than 16 million confirmed cases in 188 countries. Almost 650,000 people lost their lives in that timeframe [4].

A lot of people hope Covid-19 would be a seasonal infection. Some research supported this hope. In *Ghirelli et. al*, researchers found a significant relationship between lower temperatures and a higher effective virus reproduction number [13]. Current research also indicates that temperature, humidity, hours of sunlight, and wind speed have a direct and significant relationship with the mortality and infection rate of Covid-19 [14,15]. Experimental results from *Zohair Malki et al*. also show that weather variables are more relevant in predicting the mortality and infection rate of Covid-19 compared to demographic variables such as population, age, and urbanization [17]. Two experts in Singapore said hot weather may end the novel coronavirus [5].

However, additional research proved Covid-19 is not affected by weather. *J. Xie et. al* showed no evidence to support the theory that Covid-19 case counts decline with warmer weather [6].

Researchers in China said there's no association between Covid-19 transmission and the temperature in Chinese cities [7]. Researchers from India found the variation of temperature alone cannot explain the increased transmission of Covid-19 [8]. *Yihan Wu et. al* discovered that temperature did change people's behavior in favor of socializing more, but that did not increase the spread of Covid-19 [13]. *Stephen Afrifa et. al*. investigated climatic factors, including precipitation, temperature, relative humidity, all-sky insolation incidence on a horizontal surface, downward thermal infrared radiative flux, isolation clearness index, and wind speed at 50 meters. Their data focused on the ten countries with the highest Covid-19 cases (Spain, Italy, Turkey, Iran, the United Kingdom, the United States, Germany, France, China, and Switzerland) [19]. The researchers found that relative humidity and solar radiation have the largest influence on recorded cases of Covid-19, not temperature.

In this article, we focus on the relationship between the number of confirmed cases and the temperature in each country. First, we use the correlation coefficient to see if there is any relationship between Covid-19 and temperature. Second, we used the K-Means clustering method to divide 188 countries/regions into ten groups to determine if there is any relationship inside or between groups. Finally, we used Random Forest Regression (RFR), a machine learning technique, to find correlation or lack thereof, between temperature and Covid-19 cases. In numerous studies [16, 19, 20] RFR showed high model performance for finding the correlation between weather and increased Covid-19 cases.

## 2. DATA DESCRIPTION

We collect Covid-19 data from JHU CSSE[1]. It contains the number of cumulative confirmed cases each day in 188 countries/regions from January 22, 2020 to end of the June. Because Covid-19 was not universally present until March, we decided to use data after March 2020.

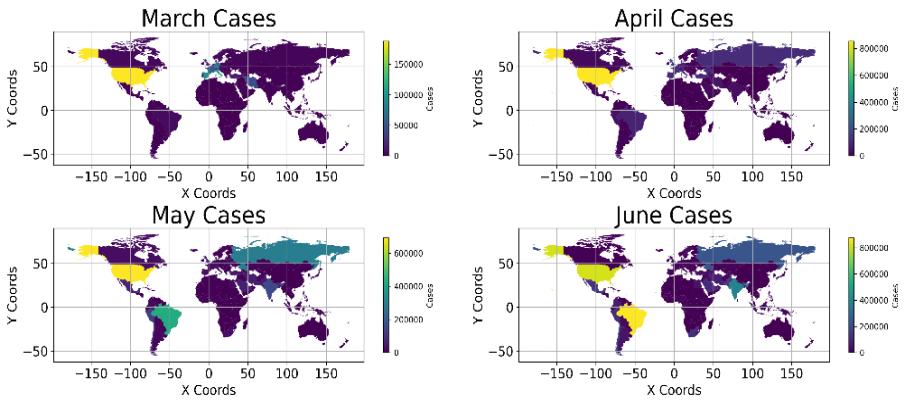



Figure 1: World maps of COVID-19 cases. A. March cases B. April cases. C. May cases. D. June cases.

Figure 1 shows confirmed COIVD-19 cases worldwide. The United States experienced the highest number of cases for the study period (March-June). Russia and Brazil did experience a significant increase in COVID-19 cases in May and June, however.

[1]https://github.com/CSSEGISandData/COVID-19

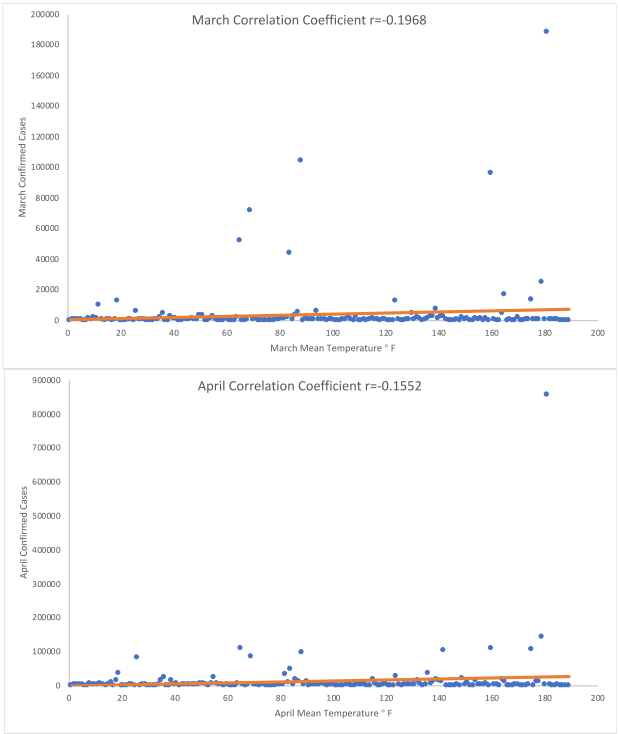
# 3. RESULTS

## 3.1. Correlation Coefficient r

The correlation coefficient of confirmed cases with average temperature is calculated monthly. The absolute value of each correlation is less than 0.3, which means there is almost no relationship between Covid-19 cases and temperature. From Figure 2, we noticed that there are some outliers, where the number of cases is more than 3 standard deviations from the mean. Many other factors, other than temperature effect confirmed cases. We also present the correlations without outliers. While the correlations are larger, they are still quite small.

Table 1. Correlation Coefficient r of Covid-19 and Month's Average Temperature

| Month | March | April | May | June |
|---|---|---|---|---|
| Pearson r | -0.1968 | -0.1551 | -0.0744 | 0.0003 |
| r After Deleting Outliers | -0.2466 | -0.2747 | -0.0079 | -0.0605 |



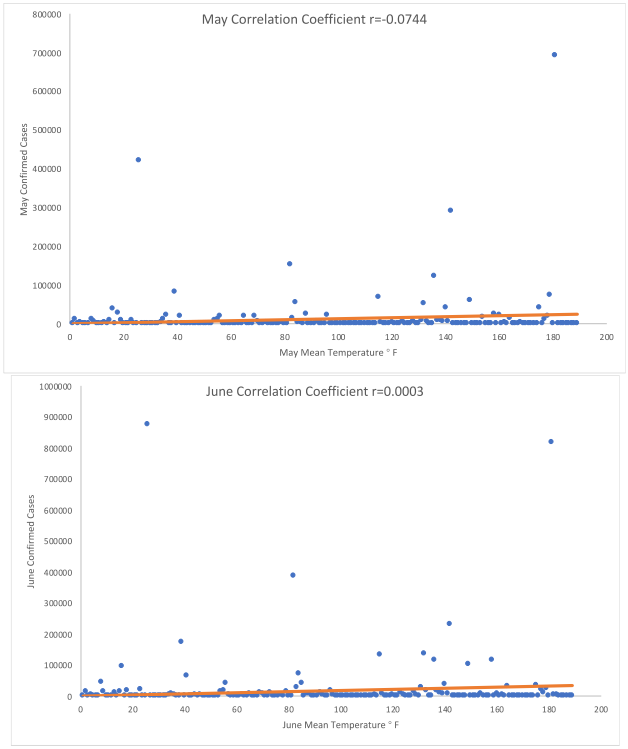



Figure 2. x-axis: Mean Temperature, y-axis: number of confirmed cases. Scatter Plots and Trend Lines between Temperature and Number of Confirmed Cases Each Month

## 3.2. K-Means Clustering

In this part we aim to cluster the countries based on the number of confirmed cases in the given months. If there is a meaningful similarity of temperature in each cluster, then we might conclude that there is a relationship between confirmed cases and temperature.

For this purpose, we used the K-means clustering algorithm. K-means is a well-established clustering method that clusters data points based on their similarity. The algorithm requires choosing the number of clusters. The elbow method is used in this analysis to find the efficient number of clusters. The elbow method uses the distortion within clusters to find the adequate number of clusters for the given data. Acceding to the elbow chart in Figure 3, 8 is an adequate number of clusters for our data.

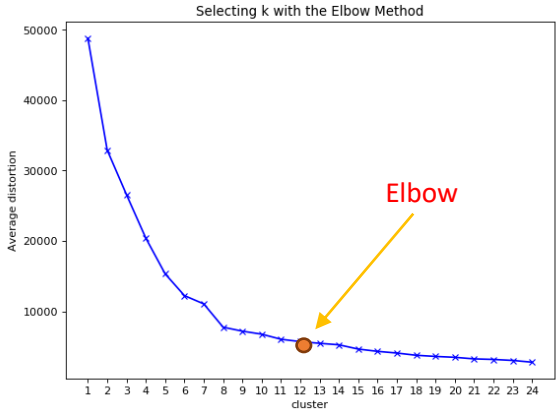



Figure 3. Selecting k with the Elbow Method

The 188 countries/regions are divided into 8 groups based on the number of confirmed cases.

Figure 4 shows the monthly temperature of the countries in each cluster. Figure 4 suggests that there is no similarity in the temperature within each cluster.

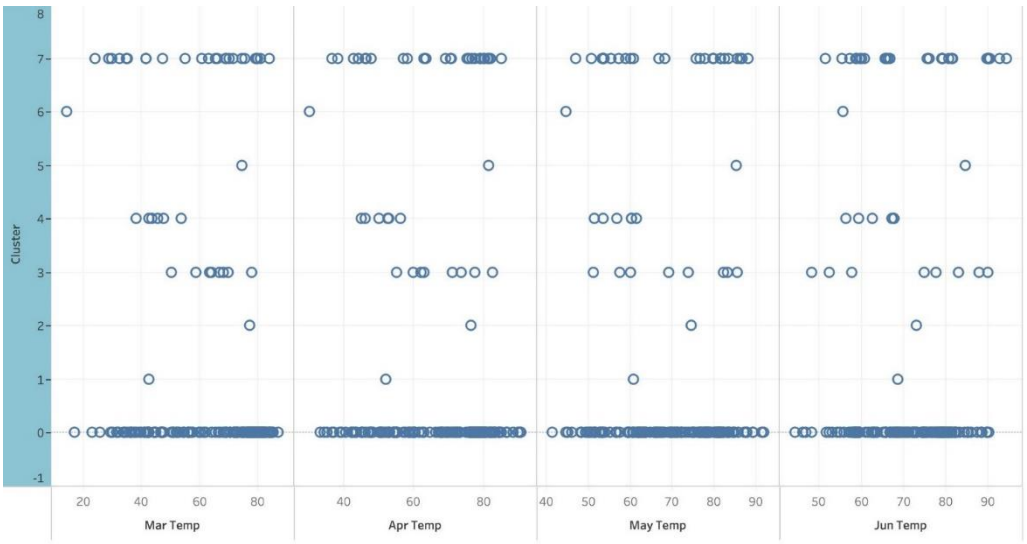



Figure 4. 8 clusters using k-mean clustering method

To illustrate the relation between the temperature and the clusters better, we use Principal Component Analysis (PCA) to project the 4 months of temperature into two dimensions. A scatter plot presents the relationship of the projected spaces regarding their clusters. Figure 5 shows the scatter plot.

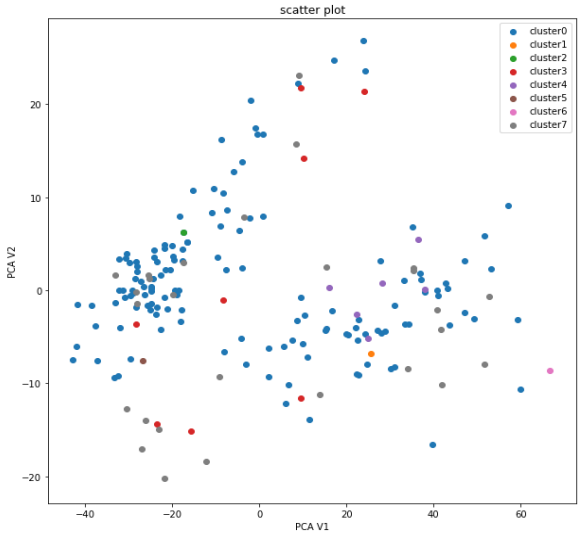



Figure 5. Scatter plot of project the 4 months temperature to 2 dimensions

According to Figure 4, there is no sensible relation between the projected temperatures (V1 and V2) and their corresponding clusters.

## 3.3. Random Forest Regression

An ensemble regressor, random forest regression, was used to predict Covid-19 cases by temperature. **Table 2** and **Figure 6** display boxplot scores for four regressors (linear regression, K-nearest neighbors, random forest regression, and support vector regression). Each regressor was trained on the Covid-19 dataset with repeated k-folds cross validation (10 data splits and15 folds). Random forest regression achieved the highest $R^2$ at 0.500618 (Table 2), therefore further research was conducted using the algorithm.

Table 2: Regression algorithms metrics.

| **Algorithm** | $\boldsymbol{R^2}$ |
|---|---|
| Linear Regression | 0.183506 |
| K-Nearest Neighbors | 0.381450 |
| Random Forest Regression | **0.500618** |
| Support Vector Regression | 0.209338 |

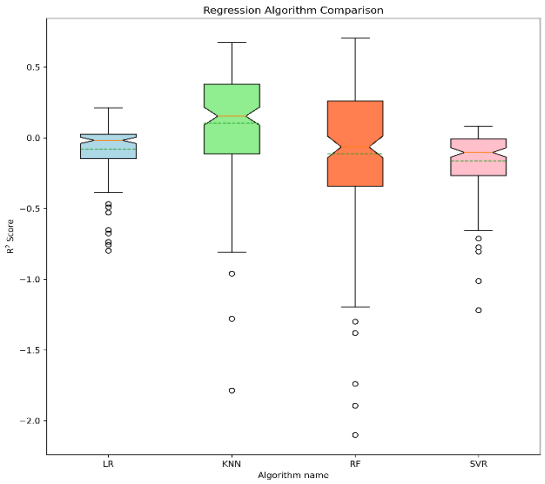



Figure 6: Boxplots of Linear Regression, K Nearest Neighbors (KNN), Random Forest Regressor (RF), and Support Vector Regressor (SVR).

An optimal number of cross validation folds is necessary to build a regression model that exhibits both low bias and low variance. MSE and SEM scores were compared for random forest regressors trained on repeated k-folds cross validation (varying the number of folds between 1 and 15). As seen in **Figure 7**, the boxplots for models March, April, May, and June show similar MSE and standard error scores after k-fold=6. The computational cost of running more cross validation folds for marginal model improvement is the rationale for picking k-fold of 7. Choosing a k-fold of 7 is a good compromise for a low bias and low variance model in the end.

With a k-folds of 7 the results are a MSE of 325.768849 and a standard error of 15.944177 for the March model; a MSE 326.773053 and a standard error of 10.994951 for the April model; a MSE 258.142170 and a standard error of 8.027189 for the May model; and a MSE 180.220601 and a standard error of 6.983111 for the June model.

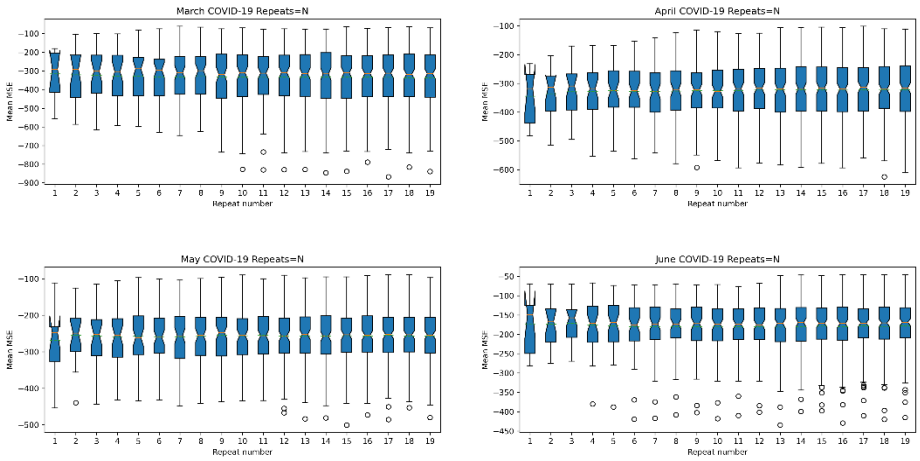



Figure 7: Boxplots of repeated k-fold cross validation (fold 1-15) Random Forest Regressor (RF).

Bayes optimization was used for hyperparameter tuning the random forest regression. This search technique optimizes model hyperparameters by using Bayes theorem to explore a search space. Compared to exhaustive search techniques like grid search, and random search techniques like randomize search, Bayes optimization only continues to explore a portion of the hyperparameter grid space when accuracy improves. The results of the random forest regression models are shown in **T Figure 8.** The March model had an MSE of 392.601, a RMSE of 19.814, a RRMSE of 30.964%, and a $R^2$ of -0.31. The April model had an MSE of 255.907, a RMSE of 15.997, a RRMSE of 23.711%, and a $R^2$ of -0.229. The May model had an MSE of 161.714, a RMSE of 12.717, a RRMSE of 18.01%, and a $R^2$ of -0.154. The June model had an MSE of 123.992, a RMSE of 11.135, a RRMSE of 15.307%, and a $R^2$ of -0.144.

| Metrics | March | April | May | June |
|---|---|---|---|---|
| MSE | 392.601 | 255.907 | 161.714 | 123.992 |
| RMSE | 19.814 | 15.997 | 12.717 | 11.135 |
| RRMSE | 30.964% | 23.711% | 18.01% | 15.307% |
| $R^2$ | -0.31 | -0.229 | -0.154 | -0.144 |

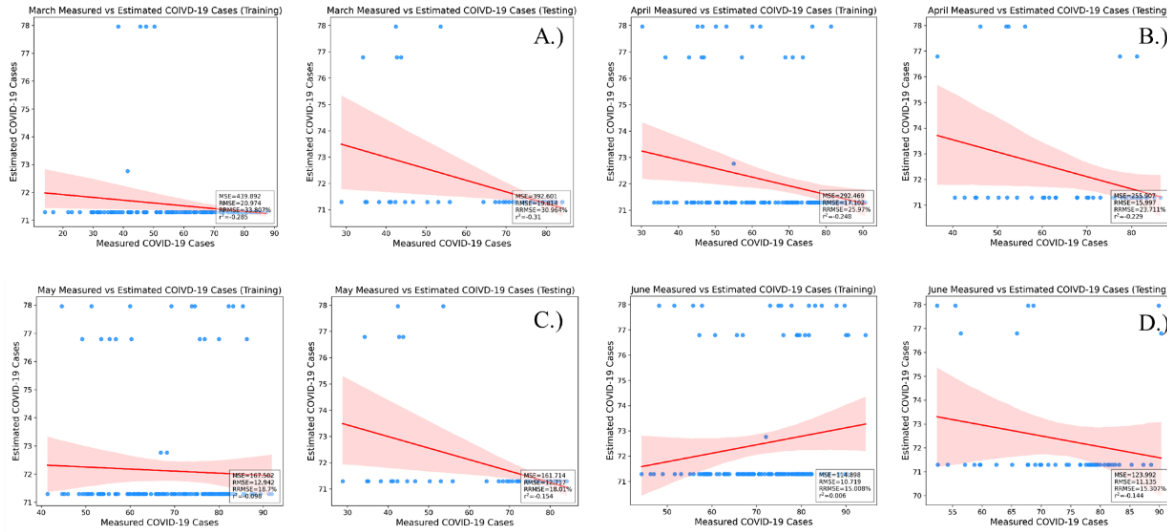



Figure 8: Scatterplots of measured COVID-19 cases versus estimated COVID-19 cases (training and testing). **A.)** March **B.)** April **C.)** May **D.)** June

The residual values for each country are displayed in **Figure 9**. The March residuals indicate that the model fits many countries in the southern hemisphere. The model does not fit well for northern hemisphere countries, however. The April residuals map shows a similar pattern with better fit in the southern hemisphere countries. One difference between the March and April model is that the April model did start to perform worse in South America. The May residuals map show a decrease in model perform for countries in the southern hemisphere, and a poor model performance in the northern hemisphere. The June residuals map shows an increase in model performance in the northern hemisphere. North America (The United States, Canada, and Mexico) specifically had very low residual values, indicating the model fit well.

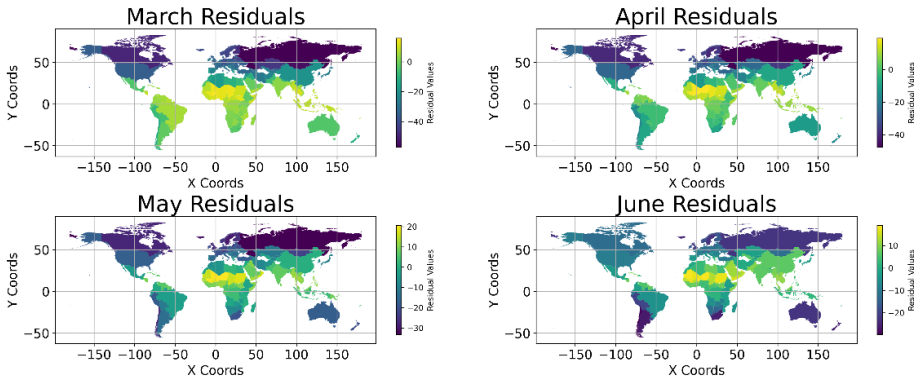



Figure 9: World maps of COVID-19 model error:March to June 2020

## 4. RESULT

Until July 28, 2020, Covid-19 was spreading worldwide, in all kinds of weather conditions. Our study above shows the number of confirmed cases has no relationship with the temperature. Random forest regressor has different fitting approaches for different months which confirms this claim. Further investigation should include a broader range of climatic data, such as humidity, wind speed, and hours of daily sunlight, to build a more accurate model for policymakers. It is also important to note that climatic data alone might not be sufficient to explain a rise in Covid-19 cases, as comfortable weather increases socializing outside, increasing the chances of spreading the virus[10].

## REFERENCES

[1] WHO: Coronavirus disease 2019 (COVID-2019) situation report-162. (2020)
[2] Tan J, Mu L, Huang J: An initial investigation of the association between the SARS outbreak and weather: with the view of the environmental temperature and its variation. J Epidemiol Community Health 2005; 59: 186–192 (2005)
[3] C. Sohrabi, Z. Alsafi, N. O'Neill: World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19), International Journal of Surgery, 71-76, (2020)
[4] HS. Badr, H. Du, M. Marshall, E. Dong, M.M.MS, L.M. Gardner: Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. The Lancet Infectious Diseases. (2020)
[5] Y. Somani, P. Tambyah: Hot and humid weather may end the novel coronavirus – as well as the development of a vaccine. (2020)
[6] J. Xie, Y. Zhu: Association between ambient temperature and COVID-19 infection in 122 cities from China, Sci Total Environ (2020)
[7] Yao Y, Pan J, Liu Z: No association of COVID-19 transmission with temperature or UV radiation in Chinese cities. Eur Respir J. (2020)
[8] Meraj, G., Farooq, M., Singh, S. K., Romshoo, S. A., Sudhanshu, Nathawat, M. S., & Kanga, S.: Coronavirus pandemic versus temperature in the context of Indian subcontinent: a preliminary statistical analysis. *Environment*, Development and Sustainability (2020)
[9] Wu Y, Jing W, Liu J: Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Science of The Total Environment (2020)
[10] Mcbryde.E, Meehan.M, O.A.Adegboye, A.I.Adekunle: Role of modelling in COVID-19 policy development. Paediatric Respiratory Reviews. (2020)
[11] Ganslmeier, M. Furceri, D., Ostry, J: The impact of weather on COVID-19. Scientific Reports. 2021.

[12] Ghirelli, C., González, A., Herrera, J.L, Hurtado, S: Weather, Mobility, and the Evolution of the COVID-19 Pandemic. Banco de Espana. 2021.
[13] Yihan Wu, Todd A. Mooringb, Marianna Linza: Policy and weather influences on mobility during the early US COVID-19 pandemic. The Proceedings of the National Academy of Sciences. 2022.
[14] A K M Bahalul Haque, Tahmid Hasan Pranto, Abdulla All Noman, Atik Mahmood: Insight about Detection, Prediction and Weather Impact of Coronavirus (COVID-19) Using Neural Network. International Journal of Artificial Intelligence and Applications, Vol. 11, No.4, July 2020.
[15] Yogesh Gupta, Ghanshyam Raghuwanshi,Abdullah Ali H. Ahmadini,Utkarsh Sharma,Amit Kumar Mishra,Wali Khan Mashwani,Pinar Goktas,Shokrya S. Alshqaq, and Oluwafemi Samson Balogun: Impact of Weather Predictions on COVID-19 Infection Rate by Using Deep Learning Models. Complexity Vol. 2021. 2021.
[16] Yasminah Alali, FouziHarrou, Ying Sun: A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. Scientific Reports 12, 2467. 2022.
[17] Zohair Malki, El-Sayed Atlam, Aboul Ella Hassanien, Guesh Dagnew, Mostafa A. Elhosseini, Ibrahim Gad: Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. National Library of Medicine: Chaos Solitons Fractals. 2020 Sep; 138: 110137. Published online 2020 Jul 17.
[18] S. T. Ogunjo, I. A. Fuwape, A. B. Rabiu: Predicting COVID-19 Cases from Atmospheric Parameters Using Machine Learning Approach. American Geophysics Union GeoHealth. 2021.
[19] Stephen Afrifa, Essien Felix Ato, Peter Appiahene, Isaac Wiafe, Rose-Mary Owusuaa Mensah Gyening, Michael Opoku: Machine Learning Impact Assessment of Climate Factors on Daily COVID-19 Cases. medRxiv. 2022.
[20] Celestine Iwendi1, Ali Kashif Bashir, Atharva Peshkar, R. Sujatha, Jyotir Moy Chatterjee, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai and Ohyun Jo: COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Front. Public Health. 2020.