Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/130216

Predicting Covid-19 Data Using Machine Learning and Statistical Methods

Abdelgalal O. I. Abaker[1], *Wahiba Ismaiel*[2], *F. M. DawAlbait*[2], *Zahra. I. Mahamoud*[3], *Hago E. M. Ali*[4], *Adil. O. Y. Mohamed*[5], *and Azhari A. Elhag*[6,*]

[1]Applied College, Khamis Mushait, King Khalid University, Abha, Saudi Arabia
[2]Department of Science and Technology, University College of Ranyah, Taif University, P.O. Box 11099, Taif 21944, Kingdom of Saudi Arabia
[3]Department of Mathematics, College of Science, Qassim University, Buraydah, 51452, Saudi Arabia
[4]Department of Business Administration, Faculty of Science and Humanity Studies, Sulail, Prince Sattam bin Adalaziz University, Al-Kharj, Saudi Arabia
[5]Department of Computer Science, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia
[6]Department of Mathematics and Statistics, College of Science, P.O. Box 11099, Taif University, Taif 21944, Saudi Arabia

Received: 22 Nov. 2023, Revised: 1 Feb. 2024, Accepted: 20 Feb. 2024
Published online: 1 Mar. 2024.

Abstract: Overall, there has been a 21% reduction in new COVID-19 cases and a 17% reduction in deaths during the most recent 28-day period (April 24 to May 21, 2023) compared to the previous 28-day period (March 27 to April 23, 2023). However, there are regional variations in the situation. The WHO Western Pacific and African Regions have seen an increase in reported cases, while the Western Pacific, African, American, South-East Asian, and American Regions have observed an increase in mortality. As of May 21, 2023, there have been a total of 6.9 million fatalities and over 766 million confirmed cases worldwide. This information can be found on the homepage of the World Health Organization (WHO). This paper focuses on the application of machine learning and statistical models to predict COVID-19 data. Both machine learning and statistical theory aim to find predictive functions from the available data through statistical inference. The study compares a time series model as a statistical approach and a decision tree model as a machine learning approach, using various statistical metrics. The findings indicate that the decision tree model exhibits the highest level of accuracy. The study specifically examines new cases of COVID-19 infection in the Kingdom of Saudi Arabia from January 3, 2020, to June 3, 2023, utilizing data obtained from the World Health Organization website.

Keywords: Machine learning, Decision tree model, Statistical model, Prediction, Forecasting Statistical inference.

1. Introduction

The Covid-19 pandemic has emerged as a global health crisis, presenting an urgent need for effective analysis and understanding of the vast amounts of data generated by the disease [1]. Traditional statistical methods have been widely used to examine Covid-19 trends and patterns, but the integration of machine learning models and time series models offers promising opportunities to derive deeper insights from this complex data [2]. Machine learning models are known for their ability to handle large and diverse datasets, identify intricate relationships, and make accurate predictions [3]. In the context of Covid-19 analysis, machine learning techniques provide a powerful framework for forecasting infection rates, estimating healthcare resource demands, and identifying influential factors [4]. Previous studies have successfully employed machine learning algorithms, such as support vector machines, random forests, and neural networks, to forecast the spread of the virus and assess its impact on various population groups [5]. Time series models, specifically designed to capture temporal dependencies, play a crucial role in understanding the dynamics of infectious diseases [6]. By leveraging the temporal nature of Covid-19 data, time series models offer valuable tools for modeling and predicting case counts, hospitalizations, and mortality rates over time [7]. Auto-regressive integrated moving average (ARIMA), exponential smoothing methods, and state space models are examples of time series models that have been applied to analyze Covid-19 data, revealing underlying patterns and trends.

This paper [8] aims to explore the effectiveness and comparative performance of machine learning models and time series models in analyzing Covid-19 data. By utilizing diverse datasets encompassing infection rates, testing data [9], and vaccination statistics, we seek to evaluate the strengths and limitations of these modeling approaches. Additionally, we will investigate the impact of different feature engineering techniques, model selection strategies, and parameter tuning

[*]Corresponding author e-mail: azhri_elhag@hotmail.com

on the predictive accuracy and interpretability of the models.

To support our analysis, we will discuss relevant studies that have utilized machine learning models and time series models in the context of Covid-19 analysis. These studies [10] have demonstrated the potential of these approaches in predicting infection rates, identifying high-risk areas, and informing public health interventions. By examining the findings and methodologies of these studies, we aim to contribute to the body of knowledge surrounding the application of machine learning and time series models in understanding and managing the Covid-19 pandemic.

Liao and Hsieh [11] focuses on predicting daily and cumulative confirmed cases of Covid-19 using statistical models. The study explores various factors influencing the spread of the virus and provides insights into forecasting Covid-19 cases.

Chinazzi et al. [12] investigate the impact of travel restrictions on controlling the spread of Covid-19. The study utilizes mathematical modeling to analyze the effectiveness of travel restrictions in mitigating the transmission of the virus.

Debnath and Berk [13] examine the potential applications of artificial intelligence (AI) in managing the Covid-19 pandemic.

The paper [14] discusses the opportunities, challenges, and ethical considerations associated with the use of AI technologies in various aspects of Covid-19 management.

Yang et al. [15] proposes a modified epidemiological model (SEIR) combined with artificial intelligence techniques to predict the trend of Covid-19 in China. The study highlights the importance of integrating mathematical modeling and AI methods for accurate prediction and effective public health interventions.

Lakhani et al. [16] focuses on time series forecasting of daily Covid-19 cases using machine learning models.

The paper [17] explores different machine learning algorithms and their performance in predicting the spread of the virus, providing insights into the potential of these models for accurate forecasting.

The paper of Mahmood et al. [18] focuses on predictive modeling and forecasting of Covid-19 using statistical learning models. It explores the application of various statistical techniques to predict the spread of the virus and provides insights into the accuracy and performance of these models.

In [19] the authors apply Bayesian hierarchical modeling to analyze the Covid-19 epidemic in the United States. The paper explores the use of hierarchical modeling to estimate parameters, assess uncertainties, and make predictions about the future course of the pandemic.

Jiang et al. [20] propose a dynamic forecasting model for Covid-19 and applies it to analyze the situation in China. The model incorporates time-varying parameters and accounts for the changing dynamics of the pandemic, providing accurate short-term forecasts and aiding decision-making.

2. Time Series Model

Time series forecasting can also be done using ARIMA models.[21] The two most popular methods for time series forecasting are exponential smoothing and ARIMA models, which offer complimentary solutions to the issue. Aiming to characterize the autocorrelations in the data, ARIMA models differ from exponential smoothing models in that they are based on a description of the trend and seasonality in the data [22]. Using a linear combination of predictors, we forecast the variable of interest in a multiple regression model. In an auto-regression model [23], the variable of interest is predicted using a linear combination of the variable's prior values. It is a regression of the variable against itself, as indicated by the word auto-regression. Alternatively, a p-order auto-regressive model [24] can be expressed as

$$Z_t = C + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \ldots.. + \theta_p y_{t-p} + \propto t \qquad (1)$$

where white noise is dented by $\propto t$, Similar to a multiple regression, except that the predictors are the lagged values of yt. it is known as an AR(p) model, which stands for an auto-regressive model of order p.

Moving average models use previous prediction mistakes in a regression-like model as opposed to prior values of the forecast variable as in a regression.

$$Z_t = C + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \ldots.. + \alpha_q \varepsilon_{t-q}, + \varepsilon_t \qquad (2)$$

White noise denoted by ε_t .then a moving average model of order q. denoted by MA(q),

Akaike's Information Criterion (AIC) [25], is used to select the predictors for regression, and it is used to determine the order of an ARIMA model. It can be written as [26]

$$\mathrm{AIC} = 2\mathrm{Log}(L) + 2(p + q + k + 1), \tag{3}$$

here L refers to the likelihood of the data.

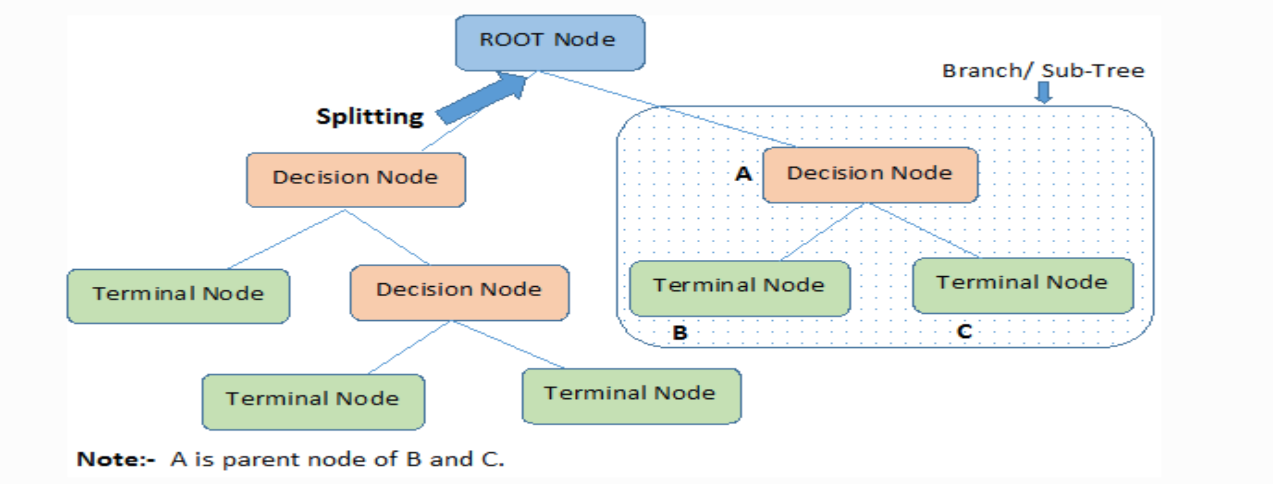

Fig. 1: Decision Tree

Decision trees are a popular classification method that organizes examples in a hierarchical structure[27], starting from the root and extending to leaf or terminal nodes, which determine the classification of each example [28]. At each node of the tree, a specific attribute is used as a test case, and the edges branching out from the node represent different potential outcomes based on the test case [29]. This recursive process is repeated for each sub-tree, allowing for further refinement of the classification. In conclusion, decision trees offer a straightforward and intuitive approach to classification tasks[30]. Their ability to capture complex decision boundaries and provide interpretability makes them a valuable tool in various domains[31]. However, careful consideration of over-fitting and appropriate parameter tuning is necessary to ensure their optimal performance.

3. Data Analysis

Daily data for new cases of Covid-19 infection in the Kingdom of Saudi Arabia were obtained from January 3, 2020, to June 3, 2023. from the World Health Organization (WHO) website.

4. Numerical results

The performance of the students that available historical in King Abdul Aziz University.

4.1 ARIMA Model for New cases

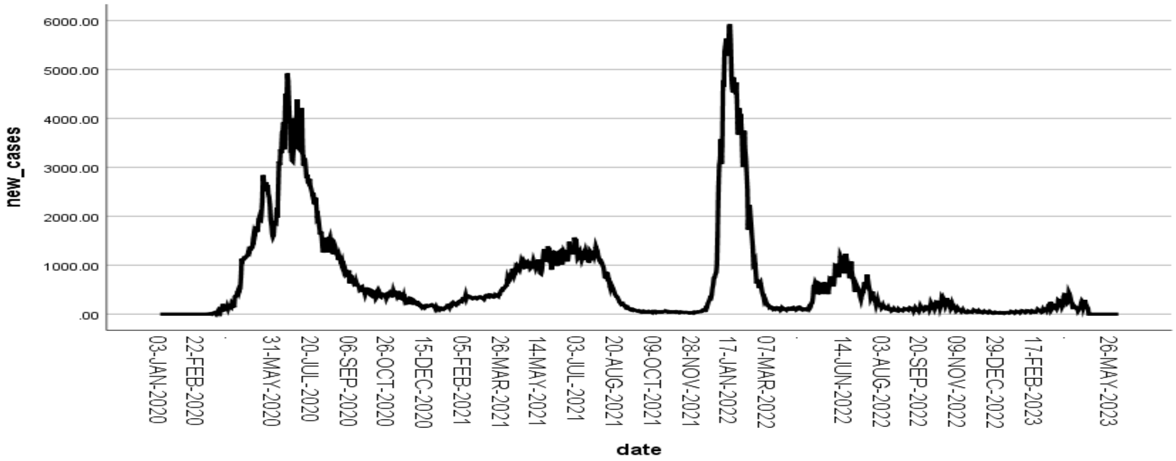

Fig. 2: Time Series Plot of New cases

The figure shows the time series of Covid-19 cases from 3.6.2020 to 3.6.2023. The figure shows two peaks for the first Covid-19 cases on 31.5.2020 and the second peak 17.5.2020.

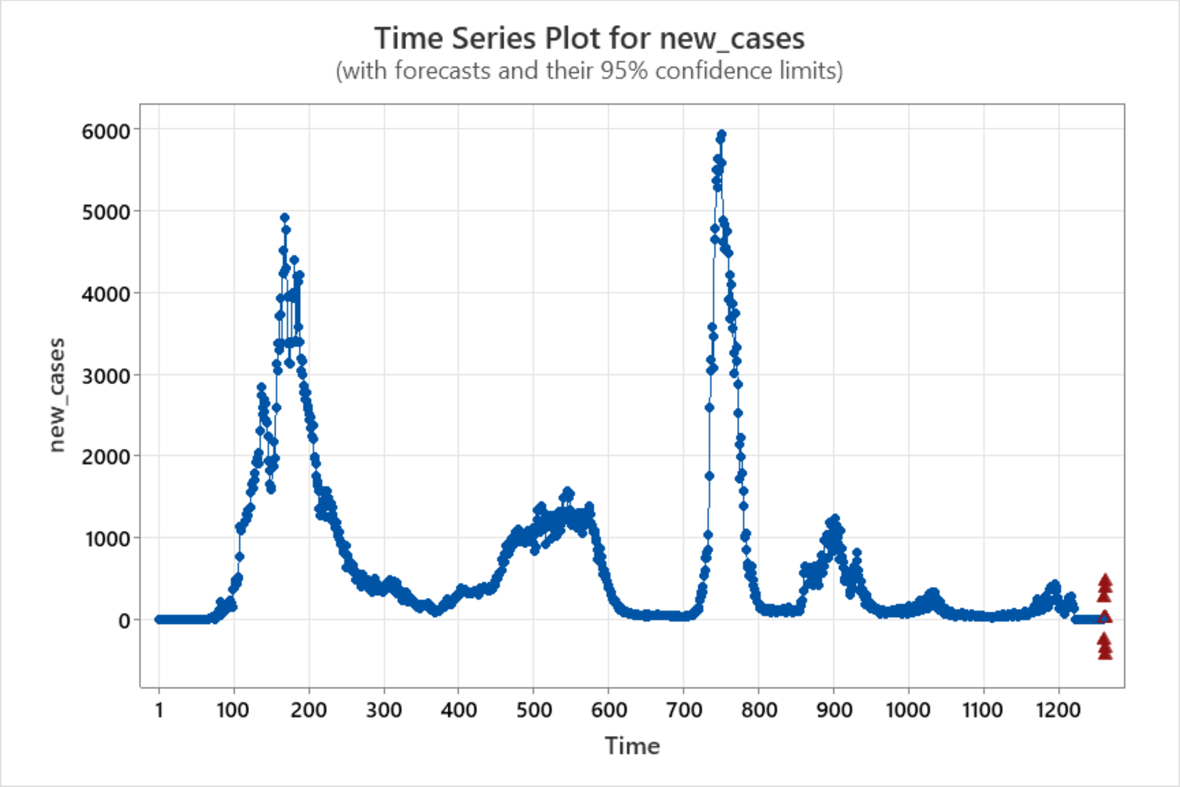

Fig. 3: Forecast with Best ARIMA Model for new_cases

Table 1: Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	-0.2812	0.0677	-4.15	0.000
AR 2	-0.6542	0.0674	-9.71	0.000
MA 1	-0.3263	0.0713	-4.57	0.000
MA 2	-0.7183	0.0725	-9.91	0.000
MA 3	-0.1880	0.0369	-5.09	0.000
MA 4	-0.0602	0.0388	-1.55	0.121
MA 5	0.0988	0.0359	2.75	0.006
Constant	-0.00	7.75	-0.00	1.000

The p-value for the auto-regressive term is below the 0.05 level of significance. The coefficient for the auto-regressive term is statistically significant, thus you may infer that the term should be kept in the model.

Table 2: The selected Model with d = 1

AR(p)	MA(q)	Loglikelihood	AICc	AIC	BIC
2	5	-7859.21	15736.6	15736.4	15782.7
1	5	-7881.42	15777.0	15778.8	15819.9
0	5	-7883.41	15780.9	15780.8	15816.8

The best model to interpret and one that is more capable of forecasting is one with fewer values. As shown from table 2. the best model is (2,1,5).

4.2 Decision Tree

Table 3: Response Information, where the mean of the response is 668, the first quartile (Q1)is 70, the median is 244 and the upper quartile (Q3)is 916.

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
668	1017	0	70	244	916	5928

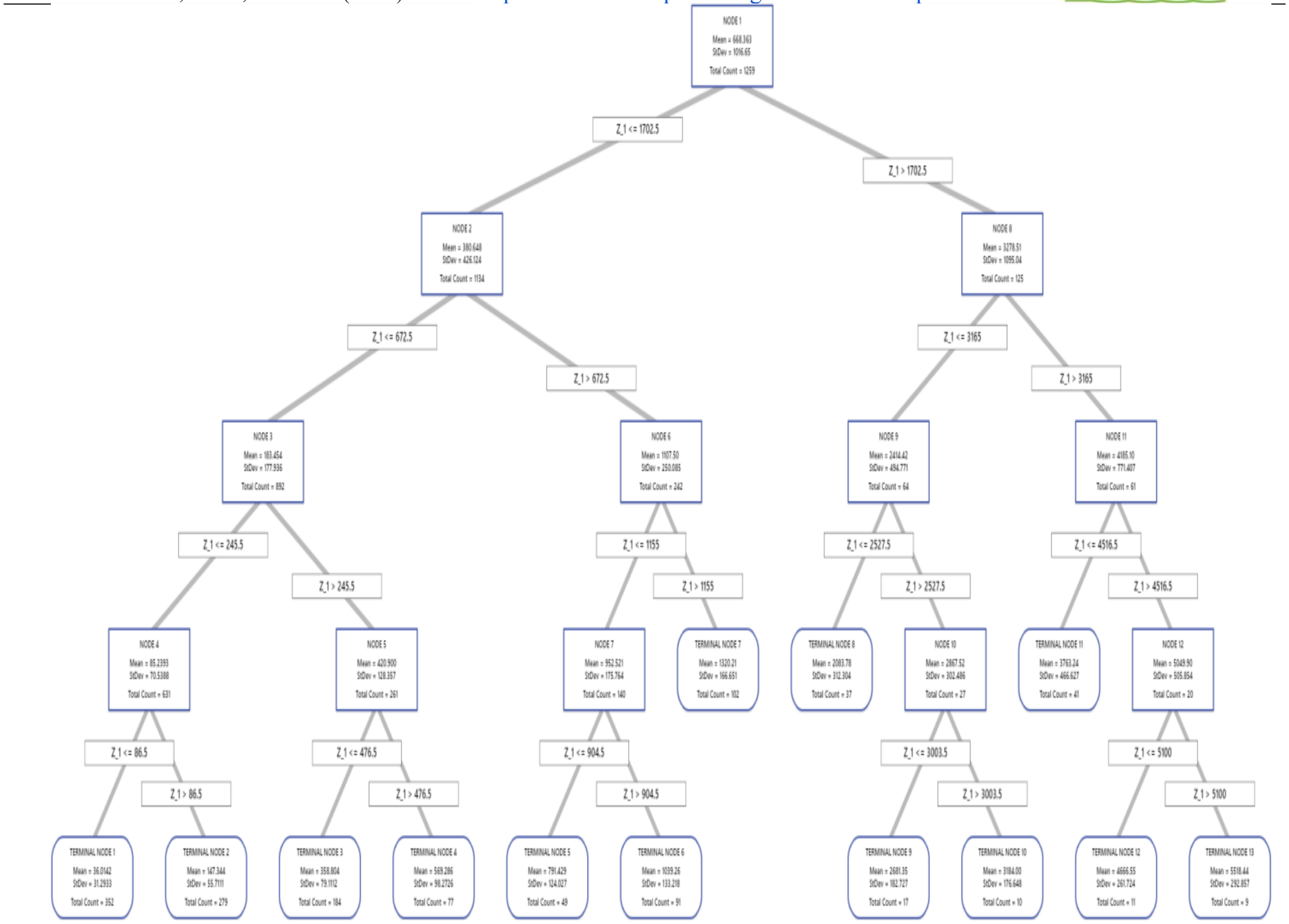

Fig. 4: Optimal Tree Diagram: The top node of the tree and the only node without parents is the root node.

Depending on the features of the data, the data at each node is divided into 2 children. Nodes split until there is no more data left in the tree to split. In other words, it is impossible to further divide the terminal nodes into groups.

Table 4: The accuracy of models : The table compares the accuracy of the training model and testing model using statistical metrics, where the lower value of MAPE is more accuracy.

	ARIMA (2,1,5)		Decision Tree	
Statistics	Training	Test	Training	Test
R-squared	98.04%	97.18%	**98.06%**	**97.61%**
Mean absolute percent error (MAPE)	0.3325	0.3383	**0.0099**	**0.0101**

5. Results and Discussion

The global impact of the Covid-19 pandemic has necessitated a shift towards learning to coexist with the virus while taking necessary precautions. Recognizing that this ongoing pandemic may persist for years to come, it becomes imperative to strike a balance between safeguarding public health and maintaining essential societal functions. In this study, our focus was on the prediction of Covid-19 cases in the Kingdom of Saudi Arabia (KSA), employing both machine learning and statistical models. Our investigation involved a comparative analysis of machine learning and statistical models to gain insights into their predictive capabilities for Covid-19. Specifically, we examined the accuracy of the decision tree model in predicting new cases of Covid-19, contrasting it with time series models based on statistical measures. The results, as presented in Table 4, highlight the superior accuracy achieved by the decision tree model. Furthermore, Figure 2 provides a visual representation of the time series data, illustrating the actual and forecast trends of Covid-19 cases. This visualization serves to enhance our understanding of the predictive performance of the models employed.

These findings underscore the potential of machine learning approaches, particularly the decision tree model, in effectively predicting new Covid-19 cases. By leveraging such models, policymakers and healthcare authorities can make informed decisions and implement targeted interventions to mitigate the spread of the virus.

6. Conclusion and perspectives

In conclusion, our study contributes to the growing body of research aimed at predicting and understanding the dynamics of Covid-19. The utilization of machine learning and statistical models provides valuable insights into the prediction of Covid-19 cases in the KSA context. As we continue to navigate this prolonged pandemic, such predictive models can play a crucial role in informing public health strategies and facilitating effective decision-making. Looking ahead, there are several promising avenues for further research and application in the field of Covid-19 prediction. Firstly, exploring the integration of real-time data sources, such as mobility patterns, social media sentiment, and environmental factors, can enhance the predictive models' accuracy and timeliness. Additionally, incorporating individual-level data, such as age, preexisting conditions, and vaccination status, can provide valuable insights into the differential impact of the virus on various population groups. Furthermore, investigating the long-term effects of Covid-19 on health outcomes and understanding the potential for recurrent waves or seasonal patterns can aid in developing proactive mitigation strategies. Collaboration between researchers, policymakers, and public health agencies is essential to harness the full potential of predictive models and ensure their effective use in guiding public health interventions and decision-making at local, national, and global levels. In summary, this study highlights the effectiveness of machine learning and statistical models in predicting Covid-19 cases. By considering these perspectives, we can further advance our understanding of the pandemic and empower decision-makers with accurate and timely information for effective response and management.

Acknowledgement:

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP2/462/44

References

[1] Tran, B. L., Chen, C. C., Tseng, W. C., & Liao, S. Y. (2020). Tourism under the early phase of COVID-19 in four APEC economies: An estimation with special focus on SARS experiences. International Journal of Environmental Research and Public Health, 17(20), 7543.

[2] de Paiva, B. B. M., Pereira, P. D., de Andrade, C. M. V., Gomes, V. M. R., Souza-Silva, M. V. R., Martins, K. P. M. P., ... & Marcolino, M. S. (2023). Potential and limitations of machine meta-learning (ensemble) methods for predicting COVID-19 mortality in a large inhospital Brazilian dataset. Scientific Reports, 13(1), 3463.

[3] Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science, 368(6489), 395-400.

[4] Tian, H., Liu, Y., Li, Y., Wu, C. H., Chen, B., Kraemer, M. U., ... & Dye, C. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science, 368(6491), 638-642.

[5] Kambouris, M. E. (2023). Global Catastrophic Biological Risks in the Post-COVID-19 World: Time to Act Is Now. *OMICS: A Journal of Integrative Biology*, *27*(4), 153-170.

[6] Yan, Q., Wang, B., Zhang, W., Luo, C., Xu, W., Xu, Z., ... & You, Z. (2020). Attention-guided deep neural network with multi-scale feature fusion for liver vessel segmentation. IEEE Journal of Biomedical and Health Informatics, 25(7), 2629-2642.

[7] Yang, Z., Zeng, Z., Wang, K., Wong, S. S., Liang, W., Zanin, M., ... & Yeung, D. S. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. Journal of Thoracic Disease, 12(3), 165-174.

[8] Gök, E. C., & Olgun, M. O. (2021). SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. Neural Computing and Applications, 33(22), 15693-15707.110121 ,140 .

[9] Syeda, H. B., Syed, M., Sexton, K. W., Syed, S., Begum, S., Syed, F., ... & Yu Jr, F. (2021). Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. JMIR medical informatics, 9(1), e23811.

[10] Nishiura, H., Linton, N. M., & Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. International journal of infectious diseases, 93, 284-286.

[11] Bouchnita, A., & Jebrane, A. (2020). A hybrid multi-scale model of COVID-19 transmission dynamics to assess the potential of non-pharmaceutical interventions. Chaos, Solitons & Fractals, 138, 109941.

[12] Lakhani, P., Prateek, G., & An, M. (2020). Time series forecasting of COVID-19 daily cases using machine learning models. Chaos, Solitons & Fractals, 140, 110186.

[13] Trivedi, S. K., Patra, P., Singh, A., Deka, P., & Srivastava, P. R. (2023). Analyzing the research trends of COVID-19 using topic modeling approach. Journal of Modelling in Management, 18(4), 1204-1227.

[14] Venna, V. R., & Ugander, J. (2017). Local graph clustering: A survey. ACM Computing Surveys, 50(3), 43.

[15] Shi, C., Qin, L., Song, G., Yuhao, K., & Xun, S. (2020). Mitigating COVID-19 outbreak via high testing capacity and strong transmission-intervention in the United States. https://www.medrxiv.org/content/medrxiv/early/2020/04/07/2020.04.03.20052720.full.pdf

[16] Wang, Q., Su, M., Zhang, M., & Li, R. (2021). Integrating digital technologies and public health to fight Covid-19 pandemic: key technologies, applications, challenges and outlook of digital healthcare. International Journal of Environmental Research and Public Health, 18(11), 6053.2070 ,(4)18 .

[17] Pun, M., Turner, R., Strapazzon, G., Brugger, H., & Swenson, E. R. (2020). Lower incidence of COVID-19 at high altitude: facts and confounders. High altitude medicine & biology, 21(3), 217-222.

[18] Iqbal, S., Qureshi, A. N., Ullah, A., Li, J., & Mahmood, T. (2022). Improving the Robustness and Quality of Biomedical CNN Models through Adaptive Hyperparameter Tuning. Applied Sciences, 12(22), 11870.

[19] Iqbal, S., Qureshi, A. N., Ullah, A., Li, J., & Mahmood, T. (2022). Improving the Robustness and Quality of Biomedical CNN Models through Adaptive Hyperparameter Tuning. Applied Sciences, 12(22), 11870.

[20] Alqahtani, M. (2023). Artificial intelligence and entrepreneurship education: a paradigm in Qatari higher education institutions after covid-19 pandemic. International Journal of Data and Network Science, 7(2), 695-706.

[21] Nazia, N., Butt, Z. A., Bedard, M. L., Tang, W. C., Sehar, H., & Law, J. (2022). Methods used in the spatial and spatiotemporal analysis of COVID-19 epidemiology: a systematic review. International Journal of Environmental Research and Public Health, 19(14), 8267.

[22] Levine, G. N., Lange, R. A., Bairey-Merz, C. N., Davidson, R. J., Jamerson, K., Mehta, P. K., ... & American Heart Association Council on Clinical Cardiology; Council on Cardiovascular and Stroke Nursing; and Council on Hypertension. (2017). Meditation and cardiovascular risk reduction: a scientific statement from the American Heart Association. Journal of the American Heart Association, 6(10), e002218.

[23] Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. Computational statistics & data analysis, 55(9), 2579-2589.

[24] Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting, 33(2), 345-358.

[25] Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. Information Sciences, 239, 142-153.

[26] Johnpaul, C. I., Prasad, M. V., Nickolas, S., & Gangadharan, G. R. (2021). Fuzzy representational structures for trend based analysis of time series clustering and classification. Knowledge-Based Systems, 222, 106991.

[27] Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological methods, 17(2), 228.

[28] Wei, X. (2021). A method of enterprise financial risk analysis and early warning based on decision tree model. Security and Communication Networks, 2021, 1-9.

[29] Sahinoglu, M. (2005). Security meter: A practical decision-tree model to quantify risk. IEEE security & privacy, 3(3), 18-24.

[30] Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

[31] Ubar, R. (1996). Test synthesis with alternative graphs. *IEEE Design & Test of Computers*, *13*(1), 48-57.