Asghar et al. LGURJCSIT 2024

LGU Research Journal of Computer Science & IT

ISSN: 2521-0122 (Online)
ISSN: 2519-7991 (Print)

doi: 10.54692/lgurjcsit.2024.082573

Vol (8): Issue (2), April – June 2024

Integrative Machine Learning Framework for Accurate COVID-19 Forecasting

Muhammad Asghar[1*], Nida Anwar[2], Manahil Hassan[3], Komal Saleem[4]

[1]Department of Computer Science and Information Technology, Virtual University of Pakistan, Lahore, Pakistan.

[2]Department of Computer Science and Information Technology, Virtual University of Pakistan, Lahore, Pakistan.

[3]Department of Computer Science and Information Technology, Virtual University of Pakistan, Lahore, Pakistan.

[4]Department of Computer Science and Information Technology, Virtual University of Pakistan, Lahore, Pakistan.

Email: ms190400021@vu.edu.pk

ABSTRACT:

This research paper introduces a machine learning (ML) forecasting model for COVID-19 cases to investigate the performance of machine learning algorithms and develop a new procedure to improve prediction efficiency. Utilizing the Multilayer Perceptron (MLP), Linear Regression (LR), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and the proposed approach, the study considers the COVID-19 data to forecast case numbers. The study contextualizes its findings through a systematic methodology of dataset compilation, algorithm interpretation, and framework development. It uses measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to evaluate the predictive performance. The trend in COVID-19 data and predictions from different algorithms is shown through graphical illustrations. The same goes for the proposed framework predictions. This study demonstrates that the proposed LR approach and the framework outperform previous MLP, KNN, and SVM models, suggesting the relevance of explainable and robust modelling solutions for COVID-19 risk assessment. Our suggested framework, which outperforms individual algorithms by averaging their results after combining them, significantly reduces prediction errors. Discussion of implications of forecasts for interventions, resource allocation, and policy decisions is done, with the need for accurate forecasting in the pandemic response highlighted. Further research can be expected to improve the framework, incorporate new machine learning (ML) techniques, and deploy real-time adaptive modelling systems. Collaboration between researchers, policymakers, and healthcare practitioners is crucial for adopting research results into practice and promoting evidence-based decision-making in outbreak response and preparedness. Generally, this study shows the progress of epidemiological forecasting and provides important information in fighting COVID-19 and future epidemics.

KEYWORDS: Covid-19 Forecasting, Machine Learning Algorithms, Neural Networks, Support Vector Machine, Linear Regression, KNN.

1. INTRODUCTION

This template is an example of formatting a paper for the Journal of Southwest Jiaotong University. The template is available online on the page for all authors on the official website of the Journal of Southwest Jiaotong University.

1.1. Background information on the COVID-19 pandemic

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has posed an unprecedented challenge to global health systems, economies, and societies since its emergence in late 2019. Initially detected in Wuhan, China, in December 2019, the virus spread rapidly across international borders.[1]

The World Health Organization (WHO) declared COVID-19 a pandemic in March 2020. Efforts to curb the spread of the virus have included large-scale testing and tracing, movement restrictions, and quarantine protocols.[2] During the pandemic, a clear understanding of the necessity for accurate prognostications of the severity of the COVID-19 spike appeared to be instrumental in organizing the subsequent public health measures, resource management, and decisions. This research aims to develop a better way of improving the COVID-19 prediction models so that a lot of effort is achieved in fighting the pandemic.

1.2. Importance of Predicting COVID-19 Cases

Prediction of COVID-19 cases has several reasons, as follows. It informs resource allocation by health authorities and policymakers, ensuring better mobilization of resources. Reliable projections enable epidemiologists and public health managers to implement effective strategies to minimize virus transmission. Forecasting also helps optimize vaccine distribution to areas with the greatest need. Furthermore, sharing projected trends with the public raises' awareness and compliance with health measures, such as self-isolation and social distancing [3].

In addition to aiding immediate pandemic response, evaluating the performance of various machine learning algorithms for COVID-19 case prediction is essential. This study systematically assesses the performance of these algorithms, proposing a new approach that leverages the strengths of multiple models to improve predictive accuracy.

1.3. Objectives and Structure

This study aims to improve forecasting tools and provide a reliable information base for evidence-based decision-making in response to COVID-19. It introduces a novel framework integrating multiple machine-learning algorithms to enhance predictive accuracy. The paper is structured as follows: The Methodology section describes the dataset and the employed machine learning algorithms, followed by a detailed explanation of the proposed framework. The Results and Discussion section presents the findings, including graphical illustrations and comparative analyses. The Conclusion summarizes the essential findings and suggests directions for future research.

By advancing the field of epidemiological forecasting, this study contributes valuable insights for combating COVID-19 and future epidemics.

2. RELATED WORK

Manuscripts should be written in English or Chinese. The title of the paper, information about the authors, abstract and keywords, and bibliography must be written in English and Chinese. You can submit an article online at the journal's website. To submit a paper, the author will first need to register. All manuscripts are peer-reviewed. The first decision is given to authors about 10-50 days after submission; acceptance for publication after revisions is done within 7-10 days (averages for articles published in this journal in the first half of 2021).

The Total Article Processing Charge (APC) is USD 600. Local VAT or Sales Tax will be added if applicable. Many national and private research funding organizations and universities explicitly cover APCs for articles from funded research projects. Waivers may be granted at the editorial office's discretion and should be discussed with the Editor when submitting the article. The editorial decision-making is decoupled from the authors' ability to pay the Processing Charges. However, authors should consider whether they have sufficient funds to cover the full APC. Journal of Southwest Jiaotong University also offers discount vouchers to selected reviewers.

Forecasting techniques are designed and implemented in various ways to avoid major disasters and make better decisions [4]. This chapter will review the literature on COVID-19 and its predictions; forecasting and its results will be described in this literature work [5]. The proposed approach aimed to predict the future course of COVID-19 using a straightforward yet innovative technique. Based on their analysis, they expect a continuous increase in confirmed COVID-19 cases, assuming the data's reliability and a continuation of the current infection pattern. They share the results of a real-time prediction exercise with important

implications for planning without any association with pledges or commitments. Their work involves forecasting the number of confirmed COVID-19 cases using powerful time-series models and examining the trend of recovered cases. The most recent predictions, which cover the period from 02 March 2020 to 21 March 2020, indicate a notable increase in the pattern of cases globally and a growth in the associated susceptibility.

Pandemic forecasting approaches have been researched thoroughly and implemented to reduce destructive effects and enhance decision-making. Now, the forecasts of COVID-19 impact contain different methods for control and prediction proposed by various authors. One of them was based on time series models aimed at predicting cases of COVID-19, and it was stated that there would be an increase in the number of confirmed cases in the future if other factors remained constant. Data on the statistics of the identified infection is reliable. The journal also had an article of Chinese origin and examined Italian and French cases to explain the possibility of managing the rate of infections to stop the virus [6].

Mathematical models like the Susceptible Exposed Infectious Removed (SEIR) model have been adjusted to include age groups and catchment areas to predict the daily new case Numbers, hospitalizations, deaths, and probable ICU bed demand with COVID-19 [7].

Other works developed artificial intelligence and machine learning models for COVID-19 prediction, namely the additive regression models, dynamic maps, MLP, VAR, and LR [8]. These studies further prove that machine learning and artificial intelligence are instrumental in combating pandemics.

Thus, a prediction model based on MLP, VAR, and LR will be developed as the machine learning method [9]. These algorithms are employed to forecast the number of COVID-19 patients in India. The author employed a free tool called WEKA and another tool called Orange to apply these models. Finally, the author analyzed the results of the three algorithms and proposed that MLP gives better outcomes than LR and VAR in terms of accuracy.

Modern machine learning and artificial intelligence techniques are elaborated to tackle the COVID-19 pandemic. The author briefly describes these techniques along with their applications in the form of a table, which helps a lot in understanding the power of machine learning, artificial intelligence, and other modern technologies. The author also explained that researchers are utilizing/practising these techniques to fight against COVID-19 more effectively [10].

In this paper, the authors have discussed different safety measures that can be taken using machine learning predictive techniques like predicting an outbreak, screening of patients, vaccine development, contact tracing, etc. It is also described that ML and AI can enhance treatment, forecasting and prediction, contact tracing, and vaccine development significantly [11].

In this paper, the researcher suggests employing a non-linear machine learning approach, specifically the Partial Derivative Regression (PDR-NML) method, for predicting the COVID-19 epidemic. An exponential second derivative linear regression model was utilized to implement this, and a computer-assisted analysis was conducted to determine the appropriate parameters within the dataset. The results of this study indicate that the proposed machine learning approach surpasses the current state-of-the-art methods in predicting COVID-19 trends within the Indian community [12]. This study suggests a unique strategy for forecasting the COVID-19 smoothed daily new cases per million. The nations' time-series data were utilized to generate GMM representations [13].

3. METHODOLOGY

3.1. Description of the Dataset

The dataset used in this research comprises a complete collection of COVID-19-related papers from sources such as national health systems, global health organizations, and research institutions. This dataset covers the epidemiological parameters and characteristics, demography, healthcare facilities & infrastructure, and socio-economic factors that are part of the countries and regions that are the epicentres of the pandemic. The key variables utilized in this study are:

- Total confirmed cases
- Total deaths
- Total recoveries
- Daily new cases
- Daily new deaths
- Testing rates

These variables contain the necessary information to study the dynamics of the pandemic and the actions taken in detail.

3.2. Explanation of Machine Learning Algorithms

In this study, four distinct machine learning algorithms are employed to predict COVID-19 cases: Among them, the most common are – Multilayer Perceptron (MLP), Linear Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM). Each algorithm introduces its properties, which qualify it to perform this task.

A. Multilayer Perceptron (MLP)

MLP is one of the feedforward neural networks used to train the data sets containing features with non-linear correlations. It is especially recommended for analyzing large data sets in which the nature of relationships is unknown.

B. Linear Regression (LR)

LR is the most basic type of algorithm where a linear equation represents the correlation between features and target variables. Due to its interpretability and applicability in situations with linear relationships can be used in COVID-19 prediction.

C. K-Nearest Neighbors (KNN)

KNN is considered one of the non-parametric methods of classification where the nearest data points directly influence the model's flexibility. This flexibility makes it quite valuable in dealing with non-linear relationship situations.

D. Support Vector Machines (SVM)

SVM is well known for its high efficiency in building the decision boundary that gives the most significant class separation. However, it is not always efficient and depends on the data's characteristics. These algorithms were chosen based on many properties so that the framework could use one strength and minimize the others' weaknesses.

Figure 1: Proposed Methodology

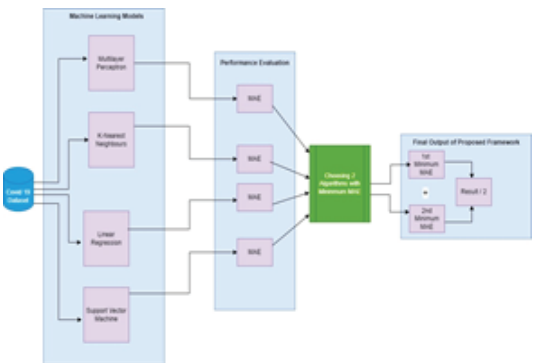

Figure 2: Our Proposed Framework

3.3. Detailed Explanation of the Proposed Framework for COVID-19 Prediction

Several Machine Learning Models are used in the COVID-19 prediction to maximize the prediction scores. This framework systematically makes predictions about the task, chooses the correct algorithm, merges the result, and assesses the performance.

A. Algorithm Selection

The four depicted algorithms, MLP, LR, KNN, and SVM, are chosen in a way that their characteristics will complement each other to improve the overall prediction ability of the model.

B. Prediction Generation

Each algorithm is used to build models and establish correspondence in the given data set with the help of epidemiological, demographic, and healthcare system indicators to forecast COVID-19 cases. With the help of iterations conducted on training and testing the models, the current models are increasingly accurate.

C. Combination of Predictions

After passing through the updating process and outlier elimination, the final forecasts are then averaged using a weighted summation model. The weights influence algorithms' accuracy levels, whereby algorithms that record high accuracy are given high weights in the final prediction.

4) Performance Evaluation:

These integrated estimations are then checked

with the help of indicators like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These give quantitative measures of the degree of prediction and measures of reliability so the framework can be compared against accurate COVID-19 data.

3.4. Mathematical Formulation

The suggested framework has been mathematically defined as the average of the minimum MAE errors of the two superior models. The formula is as follows:

$$\text{Proposed Framework}=(\text{Min1}(\sum_{i=1}^{n}|yi-y\tilde{i}|/n) + \text{Min2}(\sum_{i=1}^{n}|yi-y\tilde{i}|/n))/2 \quad (1)$$

In this formula:
Min1(∑ |yi-y ĩ| /n) is showing the minimum MAE of 1st algorithm.
Min2($\sum_{i=1}^{n}$ |yi-y ĩ| /n) is showing the minimum MAE of 2nd algorithm.
This approach takes advantage of the robust characteristics of an array of algorithms, making the prediction framework more accurate and dependable. Therefore, this methodology presents the study process, the methods to be used in this research, and the objectives and questions posed in the research study. The general outline of the research, the description of the dataset, the machine learning algorithms, and the proposed framework are pretty comprehensive and coherent, which helps in understanding all the research design and analysis procedures used in the study.

4. RESULTS AND DISCUSSION

4.1. Presentation of graphical results depicting COVID-19 data trends

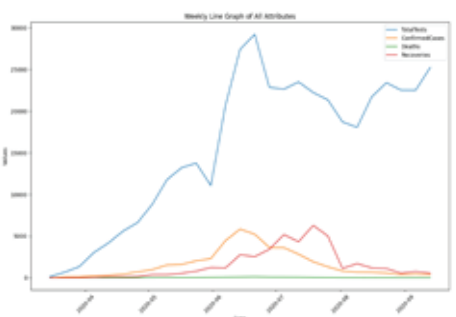

Figure 3: Weekly Graph of all Attributes

Figure 3 shows the line graph in which all dataset attributes are drawn. The highest line shows the total number of tests conducted (blue colour). Confirmed cases are shown in yellow, and recoveries are shown in red. The line at the bottom shows the number of deaths in the green line.

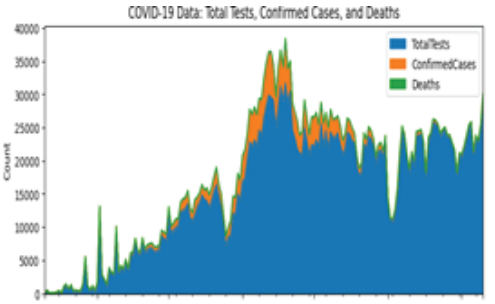

Figure 4: Stacked Graphs of all Attributes

In Figure 4, the stacked graph shows three attributes. Attributes like Total Tests, confirmed cases, and deaths are shown in different colours. Legend is also printed on the graph.

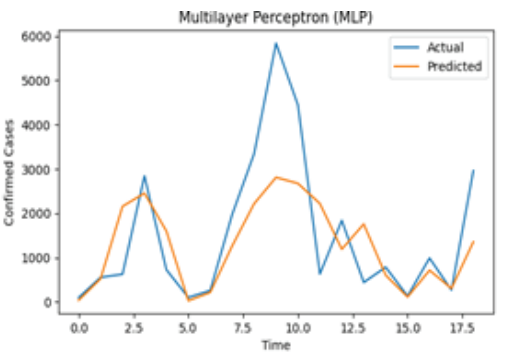

Figure 5: Prediction using MLP

In Figure 5, we can see the output of the multilayer perceptron algorithm. The actual value of confirmed cases is shown in the blue line, and the orange line shows the predicted values of confirmed cases. The X-axis shows the period. As we can see, there is an enormous gap between the blue line and orange line in the centre of the graph, so by looking at the graph, we can say that this model is not performing well.

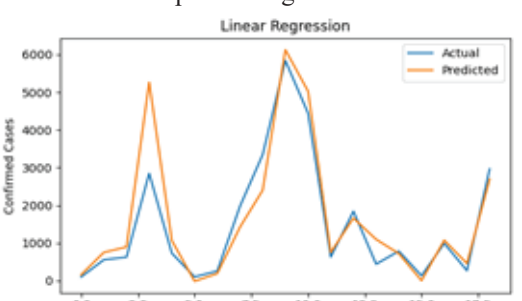

Figure 6: Prediction using Linear Regression

In Figure 6, we can see the output of the linear regression algorithm. The actual value of confirmed cases is shown in the blue line and the orange line is showing the predicted values of confirmed cases. The X-axis shows the period. As we can see both lines are very close, so by looking at the graph, we can say that this model is performing well.

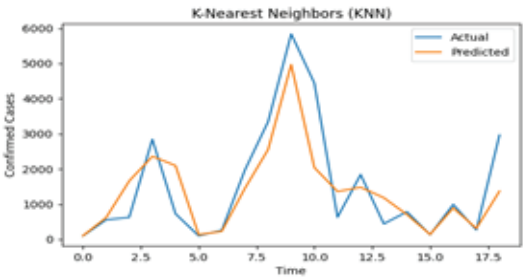

Figure 7: Prediction using KNN

In Figure 7, we can see the output of the KNN algorithm. The actual value of confirmed cases is shown in the blue line, and the orange line shows the predicted values of confirmed cases. The X-axis shows the period. As we can see, both lines are very close to each other, so by looking at the graph, we can say that this model is performing well.

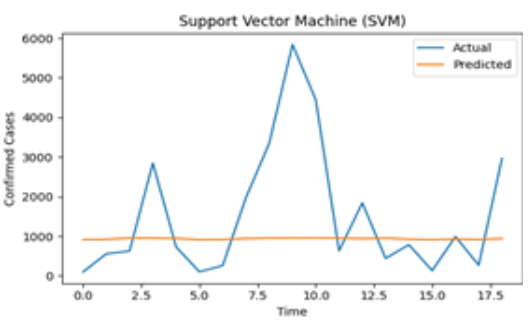

Figure 8: Prediction using SVM

In Figure 8, we can see the output of the support vector machine algorithm. The actual value of confirmed cases is shown in the blue line and the orange line is showing the predicted values of confirmed cases. The X-axis shows the period. As we can see, there is a big gap between the blue line and the orange line in the center of the graph, so by looking at the graph, we can say that this model is not performing well.

So, by analyzing all previous graphs, we can say that the two models named KNN and linear regression are performing well compared to the other two models named support vector machine and multilayer perceptron, which are not performing well. Based on our proposed model, we will develop a machine-learning framework based on the two algorithms that are performing well.

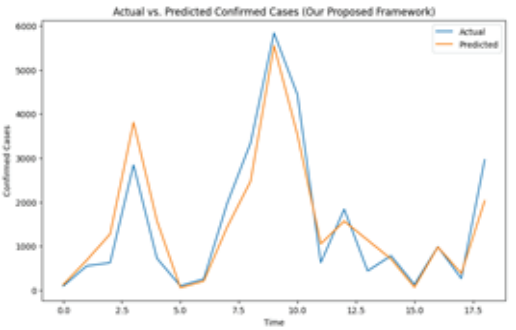

Figure 9: Prediction using our Proposed Framework

In Figure 9, we can see the output of our proposed algorithm. The actual value of confirmed cases is shown in the blue line and the orange line shows the predicted values of confirmed cases. The X-axis shows the period. As we can see both lines are very close to each other.

Table 1: Output of Different Models

Algorithm	MAE	MSE	RMSE
MLP	805.93	1300931.94	1140.58
LR	396.27	437986.66	661.80
KNN	592.29	756687.85	869.87
SVM	1172.32	2881699.18	1697.55
Our Proposed Framework	416.38	299751.13	547.49

4.2. Discussion

This section interprets the results in Table 1, providing a detailed analysis of each model's performance, its suitability for the task, and implications for COVID-19 prediction.

4.3. Analysis of Results

Results shown in Table 1 depict that the Linear Regression (LR) and also the K-Nearest Neighbors (KNN) algorithms yielded better results as compared to the Multilayer Perceptron (MLP) and the Support Vector Machine (SVM) algorithms in terms of COVID-19 case prediction. Hence, the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) of each model are the parameters that are applied as criteria for comparing their accuracy of

predictions.

A. Linear Regression (LR)

Out of all the models, LR came out on top with the least MAE of (396. 27); MSE of (437, 986. 66); and RMSE of (661. 80); hence, recommended for use in the prediction of employee attrition. Due to LR's basic structure and ease of interpretability, it is quite suitable for modelling linear association between the predictors and the response variable. Concerning COVID-19 prediction, whereby some epidemiological patterns could be linear, the ability of LR to estimate these relations is paramount to its better performance.

B. K-Nearest Neighbors (KNN)

The predictive ability of KNN was also very desirable and reasonably accurate regarding MAE of 592. 29, MSE of 756,687. 85, and RMSE of 869. 87. KNN is mainly a non-parametric model; thus, it can always strategically adjust to catch various complicated data patterns, which is useful in non-linear data. This could have boosted its performance in the role of predicting the numbers of COVID-19 cases since the patterns may fluctuate and depend on various circumstances.

C. Multilayer Perceptron (MLP)

Regarding the error metrics, MLP recorded slightly higher values than LR and KNN; thus, MAE = 805. 93, MSE = 1,300,931. 94, RMSE = 1, 140. 58. MLP excelled due to its sensitivity to the architecture and the fact that hyperparameter optimization is generally demanding. The structure, the number of layers in the neural network, and the related non-linear setting might not have been suitable for the proffered data, implying that this particular predictive task was executed with lower accuracy.

4.4. Support Vector Machine (SVM)

SVM came out as the worst-performing model with the highest MAE (1,172.32), MSE (2,881,699.18), and RMSE (1,697.55). SVM is also good at determining the classification decision boundaries but may not perform well in the case of regression tasks, and this was seen especially when it comes to COVID-19 prediction, where the data distribution, especially considering volatility, is irregular. It shows a clear decline in SVM generalizing capabilities while further demonstrating the complications and variations associated with analyzing pandemic-related data.

4.5. Performance of MLP

The following variables are known to be associated with MLP's lack of efficiency. Neural networks like MLP require significant amounts of data and careful tuning of hyperparameters. In this study, the available data and the specific characteristics of the COVID-19 dataset might not have provided the ideal conditions for MLP to excel. Additionally, the potential for overfitting and the complexity of training deep networks could have contributed to its high error metrics.

4.6. Implications and Broader Context

The findings of this study have several implications for the field of COVID-19 prediction and beyond. The superior performance of LR and KNN suggests that more straightforward, interpretable models can be highly effective for predicting pandemic trends, especially when dealing with linear or near-linear relationships. This contrasts with the often assumed necessity for complex models in all predictive tasks.

The proposed framework, which combines predictions from multiple algorithms, demonstrated the lowest error metrics (MAE: 416.38, MSE: 299,751.13, RMSE: 547.49), validating the approach of leveraging the strengths of individual models to enhance overall predictive accuracy. This ensemble method can be valuable for public health organizations, providing more reliable forecasts to inform policy and resource allocation.

5. LIMITATIONS AND FUTURE RESEARCH

As with most studies, there are some hindrances, and this work is no exception despite the positivity of the findings. The variables chosen may not include all possible factors affecting COVID-19 spread, like social interactions or policies and government measures worldwide. It is suggested that future studies investigate the inclusion of other characteristics and enhanced machine learning algorithms to attain a higher level of prediction.

In addition, the existence of real-time update systems and progressive modelling procedures is also helpful in increasing the effectiveness of the prediction models in response to the changing epidemiological status. This study highlights the importance of cooperation between academic individuals, authorities, and clinicians to implement the results acquired into practice, thus

improving the effectiveness of planning further pandemic control.

6. CONCLUSION

It is shown that the machine learning approaches, namely Linear Regression and K-nearest neighbours, can be practical for COVID-19 prediction. The proposed framework helps improve the system's accuracy with the help of many base models. The widespread findings of this study help develop public health policies and illustrate the need for building practical and understandable models to address pandemics.

Subsequent research should aim to improve and enhance the presented framework; besides that, new machine-learning approaches should be integrated, and new datasets of patients' cases should be used to enhance the accuracy of the prediction. Thus, by using such models to advance through collaborative research, it is possible to work towards figuring out how to prevent and respond to future epidemics of infectious diseases, therefore retaining the well-being of the populations in question.

REFERENCES

[1] M. Ciotti et al., "The COVID-19 pandemic," *Critical reviews in clinical laboratory sciences,* vol. 57, no. 6, pp. 365-388, 2020.

[2] B.S. Mohan and V. Nambiar, "COVID-19: an insight into SARS-CoV-2 pandemic originated at Wuhan City in Hubei Province of China," *J Infect Dis Epidemiol*, vol. 6, no. 4, pp. 146, 2020.

[3] J. Devaraj et al., "Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?," *Results in Physics,* vol. 21, pp. 103817, 2021.

[4] G. R. Shinde et al., "Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art," *SN computer science,* vol. 1, pp. 1-15, 2020.

[5] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19," *PloS one,* vol. 15, no. 3, p. e0231236, 2020.

[6] D. Fanelli, F. Piazza, Solitons, and Fractals, "Analysis and forecast of COVID-19 spreading in China, Italy and France," Chaos, Solitons & Fractals, vol. 134, pp. 109761, 2020.

[7] C. Massonnaud et al., "COVID-19: Forecasting short term hospital needs in France," *medrxiv,* pp. 2020.03. 16.20036939, 2020.

[8] A. N. Roy et al., "Prediction and spread visualization of COVID-19 pandemic using machine learning," 2020.

[9] R. A. A. Sujath et al., "A machine learning forecasting model for COVID-19 pandemic in India," vol. 34, pp. 959-972, 2020.

[10] A. Kumar et al., "A review of modern technologies for tackling COVID-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 14, no. 4, pp. 569-573, 2020.

[11] S. Lalmuanawma et al., "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," *Chaos, Solitons & Fractals,* vol. 139, pp. 110059, 2020.

[12] D. P. Kavadi et al., "Partial derivative non-linear global pandemic machine learning prediction of covid 19," *Chaos, Solitons & Fractals,* vol. 139, pp. 110056, 2020.

[13] E. Külah et al., "COVID-19 forecasting using shifted Gaussian Mixture Model with similarity-based estimation," *Expert Systems with Applications,* vol. 214, pp. 119034, 2023.