REVIEW ARTICLE ON COVID-19 OUTBREAK PREDICTION USING MACHINE LEARNING ALGORITHM

[1]Mr. M. Sakthivel, [2]Mr. C. Dinesh, [3]Mr. A. Mohaideen, [4]V. Sreetharan
[1,2,3,4]Assistant Professor, Department of Computer Science and Engineering,
[1,2,3]MallaReddy College of Engineering, Maisammaguda, Secunderabad-500 100,India
[4]St. Martin's Engineering College, Secunderabad, Telangana, India.
*Corresponding Author
E-mail: vsreetharancse@smec.ac.in

Abstract

The COVID-19 pandemic has necessitated new methods for controlling the spread of the virus, and machine learning (ML) holds promise in this regard. Our study aims to explore the latest ML algorithms utilized for COVID-19 prediction, with a focus on their potential to optimize decision-making and resource allocation during peak periods of the pandemic. Our review stands out from others as it concentrates primarily on ML methods for disease prediction.

To conduct this scoping review, we performed a Google Scholar literature search using "COVID-19," "prediction," and "machine learning" as keywords, with a custom range from 2020 to 2022. Of the 99 articles that were screened for eligibility, we selected 20 for the final review.

Our systematic literature review demonstrates that ML-powered tools can alleviate the burden on healthcare systems. These tools can analyze significant amounts of medical data and potentially improve predictive and preventive healthcare.

Keywords: COVID-19, Machine learning (ML), Prediction, Feature Selection, Artificial Intelligence (AI).

1.Introduction

The World Health Organization (WHO), declared on March 11th, 2020 the coronavirus, also known as COVID-19, a global pandemic [1]. Since then, the virus has caused widespread public health problems, costing the global economy billions of dollars in lost productivity. The existing health systems have been unable to keep up with the spread of COVID-19, underscoring the need for new methods of controlling pandemics. In response to this crisis, researchers around the world are leveraging science and technology, including medical image processing and ML [2].

The goal of ML, a branch of Artificial Intelligence (AI), is to develop systems that can learn from experience and advance without programming as in [3-5]. It can be classified into three categories: supervised learning, unsupervised learning, and reinforcement learning. In healthcare, ML is useful for processing complex and heterogeneous health data. It can be used in several ways, such as identifying and tracking outbreaks, diagnosing COVID-19, processing healthcare claims, and developing drugs and vaccines using supercomputers. The current study focuses on using ML to predict COVID-19, which is typically framed as a classification or regression problem. Some of the commonly used algorithms include Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), Gradient-boosted Decision Tree (GBDT), Extra Tree Classifier (ETC), Artificial Neural Network (ANN), Restricted Boltzmann Machine (RBM), Linear Regression (LR), Recursive Feature Elimination (RFE), and Extremely Randomized Trees (ERT). These algorithms offer various advantages, including quick classification, precise predictions, and the ability to handle complex data [6, 7].

Therefore, the motivation for studying COVID-19 prediction using ML methods is to provide valuable insights into the spread and severity of the disease, which can ultimately help save lives and mitigate the impact of the pandemic on society.

One contribution of using ML methods for COVID-19 prediction is that it enables the development of more accurate and reliable models for forecasting the spread and severity of the disease. These models can help public health officials and policymakers make informed decisions about allocating resources and implementing interventions to contain the pandemic.

Moreover, ML can help in identifying new patterns and trends in COVID-19 data, which may not be easily discernible through traditional statistical methods. This can lead to new insights into the disease and its behavior, allowing researchers to develop more effective strategies for combating the pandemic.

The paper's contribution is to spotlight the ML's ability to process and analyze large volumes of complex data quickly and accurately. This is especially important in the context of COVID-19, where there is a vast amount of data being generated from multiple sources, ML methods can help integrate and make sense of this data, facilitating the prediction of COVID-19.

2.Methodology

This scoping review looked for research on the Keywords using Boolean operators “AND/OR”. The search terms were: (“coronavirus” OR “COVID-19”)” AND (Prediction) AND (“Machine Learning” OR “ML”). All articles related to ML of COVID-19 prediction were included. The articles that are not in the English language were excluded. In addition, all articles about the virus that was not about ML were excluded. The studies were picked out of unrelated literature. Data extraction and synthesis were carried out after the duplicate articles had been located and eliminated. The Preferred Reporting Items of Systematic Reviews and Meta-Analysis (PRISMA) guidelines and its extensions for scoping reviews were followed in the completion of this work (PRISMA-ScR). The readers can have details about PRISMA-ScR in [8]. The flowchart of PRISMA shows the number of studies used. It appears in Figure 1.

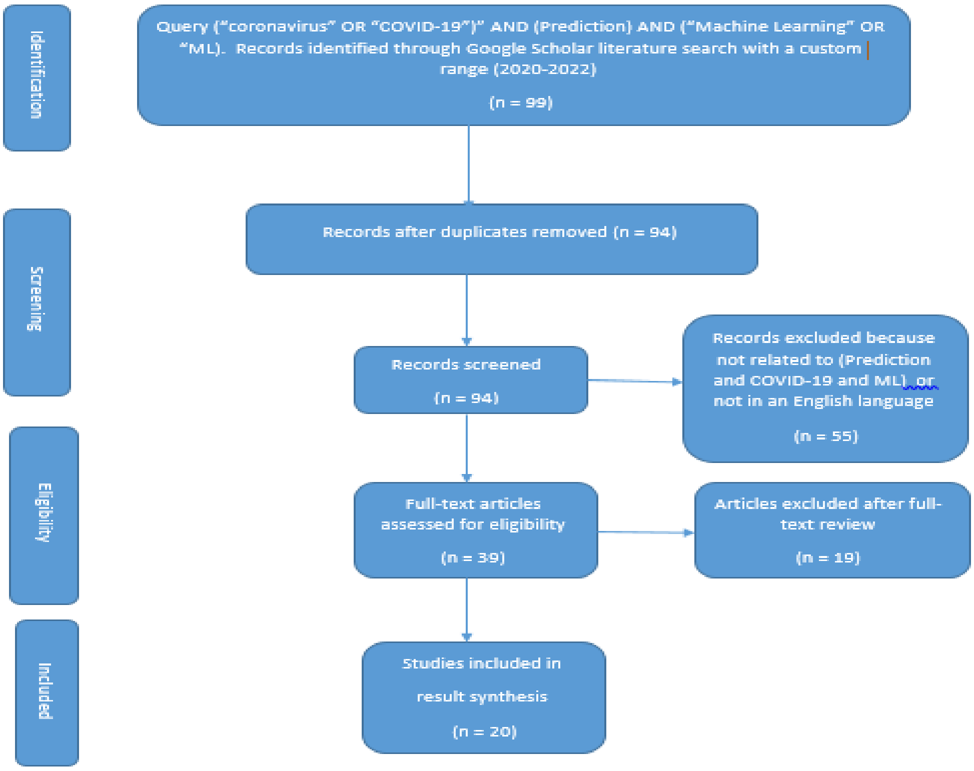
Identification
Query ("coronavirus" OR "COVID-19")" AND (Prediction) AND ("Machine Learning" OR "ML). Records identified through Google Scholar literature search with a custom range (2020-2022)
(n = 99)
Screening
Records after duplicates removed (n = 94)
Records screened
(n = 94)
Records excluded because not related to (Prediction and COVID-19 and ML) or not in an English language
(n = 55)
Eligibility
Full-text articles assessed for eligibility
(n = 39)
Articles excluded after full-text review
(n = 19)
Included
Studies included in result synthesis
(n = 20)

Figure 1. The flowchart of PRISMA.

The included studies are provided in a table form (Table 1) with their reference numbers, dataset source, the best method used, and a brief description of each study.

Table 1. Applications of modern technology to predict COVID-19 disease.

Reference number	Dataset source	Best method used	Brief description
[9]	World Bank website, Official WHO data, Weather Underground Website	RF	This study applied exponential growth models to analyze the spread of SARS-CoV-2 and determine which countries exhibited early signs of containment as of March 26, 2020. ML models were used to predict whether a pandemic would be contained within a certain time frame, and the results showed that having good healthcare infrastructure was a key factor in whether the

			pandemic was successful in being contained. The study also used four ML models (logistic regression, DT, RF, and SVM) and achieved accuracy rates ranging from 76.2% to 92.9% in predicting early containment.
[10]	Google Trends service, Official WHO Data	RF	The team developed models that can be used to predict the incidence of COVID-19 in all countries and territories worldwide. These models were accurate, with mean absolute error (MAE) ranging from 0.26 to 15.32 and Pearson correlation coefficients ranging from 0.08 to 0.99.
[11]	Johns Hopkins University, Worldometer official website, and WHO	DT	This study examined the use of ML modeling to predict the spread of the COVID-19 virus across a number of nations. According to the experimental findings, from the perspective of confirmed cases, the overall R2 is 0.99. Additionally, the model indicated when the virus would stop spreading. The findings demonstrated that the model was very successful in predicting the virus's estimated rate of spread and that it is anticipated to come to an end soon after.
[12]	The study used a labeled dataset of positive and negative COVID-19 cases from Mexico that was provided by the General Directorate of Epidemiology under the Secretariat of	DT, SVM, NB	This study used logistic regression, DT, SVM, NB, and ANN to build models for COVID-19 prediction infections in Mexico. Before developing the models, the relationship between the dependent and independent features of the dataset was analyzed using correlation coefficient analysis. The dataset was divided into 80% for training and 20% for testing the models. DT model had the highest accuracy at 94.99%, SVM had the highest sensitivity at 93.34%, and NB

	Health and had been labeled according to epidemiological criteria.		model had the highest specificity at 94.30%.
[13]	WHO	RBM	Using feature selection techniques in three stages—preprocessing, feature selection, and classification—a study has proposed a model for predicting COVID-19. The model makes use of a dataset with 8571 records and 40 features from patients in various nations. For feature selection, NB and RBM were used, and for classification, RFE and ETC were applied. The results showed that when classifying all features, the accuracy was 56.181% and 97.906% respectively, while when classifying the top ten features, the accuracy was 66.329% and 99.924% respectively.
[14]	Johns Hopkins University Center for Systems Science and Engineering dataset	ANN	This article discussed how an ML model was developed to predict the behavior of COVID-19 over a 20-day period in 30 countries. Mean squared error (MSE) and MAE were used to assess the model on a global and local level. The model was found to be effective at predicting COVID-19 behavior in each nation. Using a moving average filter with a window size of 3, the abrupt changes between samples were smoothed out and the data were normalized by the maximum value.
[15]	Johns Hopkins University's Coronavirus Resource Center	Attention-based model using Bayesian Optimizer	In this study, a combination of multi-head attention, long short-term memory (LSTM), and convolutional neural network (CNN) was utilized to create hybrid methods for COVID-19 time series forecasting. The models were designed to produce multiple predictions at different points in time, and Bayesian optimization was used

			to find the best hyper parameters for improved forecasting performance. Experiment finding showed that the proposed model outperformed the benchmark model for short-term and long-term forecasting. The best model had a mean SMAPE of 0.25 for short-term forecasting and 2.59 for long-term forecasting.
[16]	clinical information in Electronic health records (EHRs) on admission	The mortality risk prediction model for COVID-19 (MRPMC) is an ensemble model that combines four ML algorithms (LR, SVM, GBDT, and NN)	Using clinical information collected from patients at admission, a mortality risk prediction model for COVID-19 was developed. Four ML techniques (LR, SVM, GBDT, and ANN) are employed in the MRPMC model to forecast a patient's likelihood of death up to 20 days in advance. The model was validated through internal and two external validation cohorts and had an AUC of 0.9621, 0.9760, and 0.9246 respectively, This indicates that the MRPMC model has high accuracy, which can lead to more efficient and effective healthcare systems.
[17]	Ministry of Health Malaysia	Adaptive Neuro-Fuzzy Inference System (ANFIS)	The goal of the study is to use the Susceptible-Exposed-Infectious-Recovered (SEIR) model to forecast when the COVID-19 pandemic will reach its peak in Malaysia. Infection rate estimated was 0.228 ± 0.013 and takes into account cases of mortality. The pandemic's estimated peak was on July 26, 2020, with a 30-day uncertain window. If the government implements measures to decrease the infection rate by 25% over a period of two or three months, the peak could be delayed by 30 to 46 days, according to the findings of the model. The study reports that the model exhibits a low Mean Absolute Percentage Error (MAPE) of 2.45%, a low Normalized Root Mean Square Error (NRMSE) of 0.041, and a high coefficient of

			determination (R2) of 0.9964. These findings provide crucial insights for public health professionals and the government to manage the COVID-19 pandemic.
[18]	Central Hospital of Wuhan, China	LR	The goal of this study was to find a predictor of the severity of COVID-19, which is brought on by SARS-CoV-2. 377 hospitalized patients with COVID-19 were included in the study. The study evaluated and compared patients with severe and non-severe pneumonia by analyzing their clinical data, demographic information, laboratory findings, and radiological results. To identify the best predictor and validated their findings internally, the researchers used the receiver operating characteristic (ROC) curve. Age, N/L, CRP, and D-dimer were discovered to be the independent risk factors for severe pneumonia. With a cutoff value of 5.32, a product of N/LCRPD-dimer was discovered to be an effective predictor of COVID-19 severity. The study reported positive and negative predictive values of 93.75% and 46.03%, respectively, with a specificity of 89.47% and sensitivity of 67.42%. The results from the training sets indicated a negative predictive value of 93.80% and a positive predictive value of 41.32%, with a specificity and sensitivity of 70.76% and 89.87%, respectively.
[19]	Worldometer website	multi-layered perceptron (MLP)	In place of the conventional susceptible-infected-recovered (SIR) and susceptible-exposed-infectious-removed (SEIR) models, this study evaluated the efficacy of ML and soft computing models in predicting the COVID-19 outbreak. Finding

			indicated that two ML models, namely MLP and ANFIS, showed promising performance. The paper suggests that ML can be an effective tool for modeling the outbreak and recommends integrating ML and SEIR models for more accurate predictions in the future.
[20]	The national and provincial health commissions and data for natural language processing (NLP) are obtained from dxy.com, social media, and news media.	Hybrid AI	The article presented a new hybrid AI model for predicting COVID-19, called the susceptible-infected model, which outperforms traditional models by considering varied infection rates and incorporating NLP and LSTM to account for prevention measures and public awareness. Testing on data from Chinese cities revealed that COVID-19 has higher transmission rates between the third and eighth day after infection, which is more accurate than traditional models. The hybrid model also showed lower prediction errors, with a range of 0.05% to 0.86% for the next six days in various cities and nationwide.
[21]	An application programming interface	support vector regression (SVR)	With a focus on one, three, and six-day forecasts, this study assesses the precision of six ML models for predicting COVID-19 cumulative confirmed cases in ten Brazilian states. A stacking-ensemble method using cubist regression (CUBIST), RF, RIDGE, and SVR as base-learners and Gaussian process as a meta-learner are among the models, along with autoregressive integrated moving average (ARIMA), CUBIST, ridge regression (RIDGE), and SVR. SVR and stacking-ensemble models typically outperformed other models when improvement index, MAE, and symmetric mean absolute percentage error were used to evaluate the

			models. The models provided accurate forecasting with errors ranging from 0.87% to 6.90%. SVR outperformed the other models by a significant margin.
[22]	Worldometer website and Google Trends	LR	This study used LR and LSTM models to estimate the number of positive cases in an effort to predict the spread of COVID-19 in Iran. The models were tested ten times with cross-validation, and root means square error (RMSE) was used to calculate the outcomes. The frequency of searches for topics relating to handwashing, hand sanitizers, and antiseptics was the most effective variable according to the LR model, which had an RMSE of 7.562 (SD 6.492). The LSTM model had an RMSE of 27.187 (SD 20.705).
[23]	Blood samples from 404 infected patients in the region of Wuhan, China	XGBoost	This study aimed to develop a prognostic model for COVID-19 using three features: LDH, hs-CRP, and lymphocytes. The authors used XGBoost ML to create a model that can predict the survival rates of severe COVID-19 patients with an accuracy of over 90%. This model was tested on a sample of 404 patients and showed 90% accuracy for other blood samples as well. This model could help with early detection, and intervention, and potentially reduce mortality in high-risk COVID-19 patients.
[24]	WHO daily situation reports	Fuzzy Time Series (FTS)	This article discusses how accurate forecasting can help government officials make better decisions about how to spread a pandemic. Data from the COVID-19 database in India was examined from 17 March to 1 July 2020. Two models, an ANN and an FTS, were compared. The FTS model

			was found to be more precise than the ANN model in forecasting the new cases and new deaths time series. They also found that there was a short-term trend in the spread of the virus according to the models. This information can help government officials make better preparations for the health system in India.
[25]	USA Facts website	ANN	In order to forecast COVID-19 incidence rates across the continental United States, they gathered information on 57 potential explanatory variables and used an MLP. The study found that a single-hidden-layer MLP can account for nearly 65% of the relationship between incidence rates and ground truth. The study analyzed various factors that could affect COVID-19 incidence rates, and found that socioeconomic and environmental factors, including median household income and total precipitation, were important predictors. The study also used sensitivity analysis and logistic regression to determine that these factors may also contribute to the presence of hotspots in disease incidence. These findings underscore the significance of socioeconomic and environmental factors in relation to COVID-19 incidence rates.
[26]	Worldometer website	MLP Imperialist Competitive Algorithm (MLP-ICA)	The COVID-19 pandemic is predicted using a hybrid ML approach in this study. By the end of May, both the outbreak and the overall mortality are expected to significantly decline, according to the hybrid ML methods of the MLP-ICA models and adaptive network-based fuzzy inference system (ANFIS). The model demonstrated its accuracy during the nine-day validation

			process, providing encouraging results. If there are no significant disruptions, the model is expected to maintain its accuracy.
[27]	Johns Hopkins University Center for Systems Science and Engineering dataset	Exponential Smoothing (ES)	Four ML models (LR, LASSO, SVM, and ES) were found to be effective in predicting the number of COVID-19 cases in this study. The models were tested using a dataset containing information on the virus's characteristics, such as the number of new cases, fatalities, and recoveries, over the next 10 days. According to the findings, the ES model had the best ML performance, followed by the LR and the LASSO algorithm. The SVM model performed poorly in all the scenarios.
[28]	The "Daily Technical Report" issued by the Ministry of Health in Mexico	ANN	This paper uses mathematical (Gompertz and Logistic and computational models (ANN) to predict the number of COVID-19 infection cases in Mexico, based on only confirmed cases as of May 8th. The observed data and predictions had a good fit with R2 values of 0.9998, 0.9996, and 0.9999 for three models. Using these models, the number of predicted COVID-19 cases from May 9th to May 16th was 47,576, and the predicted total number of cases until the end of the epidemic was 469,917, 59,470, and 70,714 for each of the models.

3.Discussion

A prediction model is a tool used by decision-makers in healthcare to adopt an evidence-based approach and reduce mortality, and economic losses at various levels, but the accuracy and reliability of these models are not always good. Although mathematical or statistical models typically have adequate reliability, they perform poorly when faced with large amounts of data or complex phenomena [29].

ML techniques have been widely employed to create prediction models for COVID-19 due to their potential in handling the complexities of decision-making and data analysis during the pandemic. Most of the existing ML applications for COVID-19 prediction are based on supervised learning and are typically classified as classification problems, however, they have not yet been implemented in real-world settings. The ML models that have been developed so far, however, have shown promising results in predicting COVID-19. The 20 articles reviewed used different methods to predict disease and supervised learning was commonly used. The studies reviewed also used different methods to evaluate their outcomes, but most used accuracy for classification problems or R2 for regression problems. The best-obtained accuracy was 99.924% using RBM in [13], while the best-obtained R2 was 0.9999 using ANN in [28].

Early prediction can help identify high-risk individuals and provide them with the necessary care, reducing the strain on healthcare systems and allowing for more efficient resource allocation. Feature selection is used to increase the accuracy and interpretability of prediction models by only considering the most relevant features using filter methods, wrapper methods, or hybrid feature selection which combines both filter and wrapper methods with feature selection within learning algorithms [30].

The lack of extensive clinical data poses a challenge to controlling the spread of COVID-19. As a result, balancing the importance of data privacy with public health is crucial, and close collaboration between ML and humans is necessary [6].

4. Conclusion

This study demonstrates the promising value of ML in predictions of COVID-19 and the important role it can play in preparing healthcare systems so they can avoid collapse. This can be beneficial in many ways, such as planning, mitigating, diagnosing, and treating patients. This means that the future of healthcare lies with ML-based methods.

5. References

[1] F. S. Saleem, and I. Al-Mejibli, “SMARTPHONE APPLICATIONS FOR CONTACT TRACING IN COVID-19 EPIDEMIC: A SYSTEMATIC REVIEW,” *Journal of Al-Qadisiyah for computer science and mathematics,* vol. 13, no. 2, pp. Page 37–45-Page 37–45, 2021.

[2] M. Naseem *et al.*, “Exploring the potential of artificial intelligence and machine learning to combat COVID-19 and existing opportunities for LMIC: a scoping review,” *Journal of Primary Care & Community Health,* vol. 11, pp. 2150132720963634, 2020.

[3] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Speech Emotion Recognition Using Minimum Extracted Features." pp. 58-61, 2019.

[4] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, “Emotion recognition system based on hybrid techniques,” *International Journal of Machine Learning and Computing,* vol. 9, no. 4, 2019.

[5] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, “Facial emotion recognition from videos using deep convolutional neural networks,” *International Journal of Machine Learning and Computing,* vol. 9, no. 1, pp. 14-19, 2019.

[6] N. Alballa, and I. Al-Turaiki, “Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review,” *Informatics in medicine unlocked,* vol. 24, pp. 100564, 2021.

[7] S. Ustebay *et al.*, “A comparison of machine learning algorithms in predicting COVID-19 prognostics,” *Internal and Emergency Medicine*, pp. 1-11, 2022.

[8] D. Moher *et al.*, “Reprint—preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement,” *Physical therapy,* vol. 89, no. 9, pp. 873-880, 2009.

[9] D. Kasilingam *et al.*, “Exploring the growth of COVID-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning,” *Transboundary and Emerging Diseases,* vol. 68, no. 3, pp. 1001-1018, 2021.

[10] Y. Peng *et al.*, “Real-time prediction of the daily incidence of COVID-19 in 215 countries and territories using machine learning: model development and validation,” *Journal of Medical Internet Research,* vol. 23, no. 6, pp. e24285, 2021.

[11] Z. Malki *et al.*, “The COVID-19 pandemic: prediction study based on machine learning models,” *Environmental science and pollution research,* vol. 28, pp. 40496-40506, 2021.

[12] L. Muhammad *et al.*, “Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset,” *SN computer science,* vol. 2, pp. 1-13, 2021.

[13] R. H. Ali, and W. H. Abdulsalam, “The Prediction of COVID 19 Disease Using Feature Selection Techniques,” *Journal of Physics: Conference Series,* vol. 1879, no. 2, pp. 022083, 2021.

[14] P. H. Borghi, O. Zakordonets, and J. P. Teixeira, “A COVID-19 time series forecasting model based on MLP ANN,” *Procedia Computer Science,* vol. 181, pp. 940-947, 2021.

[15] H. Abbasimehr, and R. Paki, “Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization,” *Chaos, Solitons & Fractals,* vol. 142, pp. 110511, 2021.

[16] Y. Gao *et al.*, “Machine learning based early warning system enables accurate mortality risk prediction for COVID-19,” *Nature communications,* vol. 11, no. 1, pp. 5033, 2020.

[17] A. Alsayed *et al.*, “Prediction of epidemic peak and infected cases for COVID-19 disease in Malaysia, 2020,” *International journal of environmental research and public health,* vol. 17, no. 11, pp. 4076, 2020.

[18] Y. Zhou *et al.*, “A new predictor of disease severity in patients with COVID-19 in Wuhan, China,” *MedRxiv*, pp. 2020.03. 24.20042119, 2020.

[19] S. F. Ardabili *et al.*, “Covid-19 outbreak prediction with machine learning,” *Algorithms,* vol. 13, no. 10, pp. 249, 2020.

[20] N. Zheng *et al.*, “Predicting COVID-19 in China using hybrid AI model,” *IEEE transactions on cybernetics,* vol. 50, no. 7, pp. 2891-2904, 2020.

[21] M. H. D. M. Ribeiro *et al.*, “Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil,” *Chaos, Solitons & Fractals,* vol. 135, pp. 109853, 2020.

[22] S. M. Ayyoubzadeh *et al.*, “Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study,” *JMIR public health and surveillance,* vol. 6, no. 2, pp. e18828, 2020.

[23] L. Yan *et al.*, “A machine learning-based model for survival prediction in patients with severe COVID-19 infection,” 2020.

[24] P. Mishra *et al.*, “Trajectory of COVID-19 data in India: investigation and project using artificial neural network, fuzzy time series and ARIMA models,” *Annual Research & Review in Biology*, pp. 46-54, 2020.

[25] A. Mollalo, K. M. Rivera, and B. Vahedi, “Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States,” *International journal of environmental research and public health,* vol. 17, no. 12, pp. 4204, 2020.

[26] G. Pinter *et al.*, “COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach,” *Mathematics,* vol. 8, no. 6, pp. 890, 2020.

[27] F. Rustam *et al.*, “COVID-19 future forecasting using supervised machine learning models,” *IEEE access,* vol. 8, pp. 101489-101499, 2020.

[28] O. Torrealba-Rodriguez, R. Conde-Gutiérrez, and A. Hernández-Javier, “Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models,” *Chaos, Solitons & Fractals,* vol. 138, pp. 109946, 2020.

[29] M. Jamshidi *et al.*, “Hybrid deep learning techniques for predicting complex phenomena: A review on COVID-19,” *AI,* vol. 3, no. 2, pp. 416-433, 2022.

[30] J. Gong *et al.*, “A tool for early prediction of severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China,” *Clinical infectious diseases,* vol. 71, no. 15, pp. 833-840, 2020.