HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 2 (2024)

COVID-19 Forecasting Using Recurrent Neural Network and Machine Learning

Aahan Shah[1] and Kieu Pham[#]

[1]Milpitas High School, USA
[#]Advisor

ABSTRACT

The COVID-19 variant's complexity and the dire consequences of the variant spread are the inspirations behind the profound research on modeling and predicting new emerging variant surges. Multiple factors, including the variant characteristics, vaccination rate, the immune response of vaccinated individuals, and disease prevention health policies, impact the COVID-19 infection trend. The advancements in machine learning and neural network models, combined with the growth in computing, have demonstrated outstanding potential in modeling and predicting epidemic diseases. This research presents the modeling and prediction of the combined COVID-19 variant infection trends using the Holt-Winters exponential smoothing and seasonal auto-regressive integrated moving average with exogenous factors (SARIMAX) time-series machine learning models and recurrent neural network (RNN) long short-term memory (LSTM) model. Real-world United States COVID variant data from the Centers for Disease Control and Prevention (CDC) is used for prediction. The SARIMAX model factoring the seasonality of COVID-19 infections showed higher prediction accuracy than the Holt-Winters model, which is heavily weighted towards the most recent trends. The LSTM model had the best prediction accuracy of 91% with the lowest root mean square error (RMSE) values due to its property of selectively remembering patterns for long duration and the forget gates that correct the vanishing gradient problem, minimizing the error losses. This research demonstrates the promising application of the neural network deep learning models for epidemic disease modeling and prediction, enabling timely assessment of different policy decisions to mitigate the impact of an epidemic.

Introduction

The COVID-19 disease has severely impacted livelihood, economy, and people's well-being worldwide over the past few years. COVID-19 and its variants have caused more than 6 million hospitalizations and have claimed more than 1 million lives in the United States, according to data tracked by the Centers for Disease Control and Prevention (CDC), and over 6.5 million deaths worldwide, according to the World Health Organization. Studies conducted in 2020 showed that COVID-related deaths were significantly higher among persons aged 65 years or older and from minority racial and ethnic groups (Wortham et al., 2020; Gold et al., 2020). COVID-19 has many mutations that can proliferate, making it more deadly and unpredictable. SARSCoV-2, the virus that causes COVID-19, continues accumulating mutations in its genetic code, and new variants are expected to evolve. The recent Omicron variant was more transmissible than its predecessors and quickly dominated the infections (Kannan et al., 2021; Shrestha et al., 2022). The COVID variant's complexity, lack of understanding of its catalysts, and dire consequences of the variant spread, including its disproportionate impact on the aged and ethnic minorities, are the inspirations behind the profound research on modeling and predicting new emerging variant surges. A reasonably accurate infection forecasting model of the combined emerging COVID variants is required.

Over three years of COVID data from millions of people is available from the CDC data tracker to build prediction models of emerging variant surges. While good historical data is available, COVID infection modeling and prediction is a daunting task due to the multiple variants contributing to the infections at any given point in time, the complex nature of the variants' characteristics, such as infection rate and duration combined with multiple external factors such as preventative care measures, health care resources, and demographics. Since the inception of the COVID-19 pandemic, various research using different modeling techniques and approaches have been conducted to model and predict COVID-19 infection trends. Polo et al. (2020) pointed out the challenges with epidemiological disease prediction using classical surveillance approaches and recommended an approach for implementing wastewater-based epidemiology prediction. Zou et al. (2020) presented a novel epidemic model by considering the untested/unreported cases trained by machine learning models for forecasting COVID-19 spread. Alali et al. (2022) developed a Bayesian optimized Gaussian process regression (OGPR) model applied to the historical COVID cases in India and Brazil, which showed lower prediction errors compared to other static and dynamic models, such as Boosted Trees, Decision Trees, Random Forests, and XGBoost. Narin et al. (2021) trained convolutional neural network ResNet and Inception deep learning models on pneumonia-infected patients' X-ray radiographs, attaining a high COVID-19 disease prediction accuracy.

Multiple research studies have been conducted since the inception of the COVID pandemic to predict infections for a specific population in a particular region or a variant using statistical or machine learning techniques. To the author's knowledge, there is a lack of prediction techniques to forecast the combined variants' COVID-19 infection trend despite good historical COVID data availability and advances in mathematical modeling, artificial intelligence, and machine learning. COVID-19 infections follow a time series where the infections happen over time. This research presents advanced time series modeling techniques that are highly efficient in forecasting COVID-19 combined variant infections with acceptable accuracy. Holt-Winters exponential smoothing and seasonal auto-regressive integrated moving average with exogenous factors (SARIMAX) models that are better for short-term predictions (Omane-Adjepong et al., 2013) and the recurrent neural network (RNN) long short-term memory (LSTM) model that has shown to be far superior to conventional regression-based models (Siami-Namini et al., 2019) will be evaluated and compared for best prediction outcomes. The prediction model, with acceptable prediction accuracy and simplistic application, can assist proactive policy decision changes around vaccination forecasting, hospitalization resources, and medical equipment needs and ensure a swift response to an infection surge to mitigate the exposure of an epidemic.

Method

Holt-Winters and SARIMAX time-series machine learning models and LSTM RNN model are trained and tested on the publicly available US COVID variants aggregated cases data tracked by the CDC to model and predict infection cases. Various tests were run on the data to select the best model parameters to attain the highest prediction accuracy. The model prediction results were compared using the standard mean absolute error (MAE) and root mean square error (RMSE) metrics. The best prediction model obtained from this experiment can be integrated with the case tracking system to perform real-time predictions for the entire population or a subset to understand the impacts of any new emerging variant and actions required to mitigate the infection growth. The model parameters can be tuned in real-time as it learns from the new variant behaviors to improve the prediction accuracy over time.

Experimental Design

The COVID-19 infection modeling design included data acquisition, data processing, cleaning and preparation, model selection, coding and model development, and prediction assessments, as shown in Figure 1. Real-world

US COVID-19 variant infection case data were extracted from the CDC tracking database. The models were trained and tested on 21 months of COVID variants aggregated case data. The COVID-19 cases data were processed, grouped by week, and trended over time.

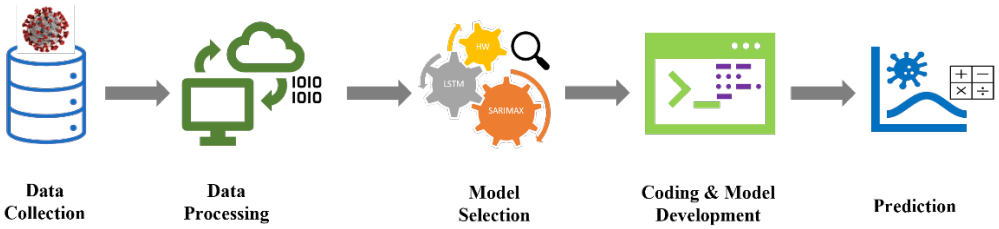

Figure 1. Experimental design steps.

Time-series machine learning and RNN models were selected for COVID-19 infection case modeling and prediction for their versatility, simplicity, and applicability. Time series analysis can be performed on the COVID-19 case data, as it can be grouped and ordered in time. Time series can generally be decomposed into three components: trend, seasonality, and noise. The upward trend shows the series is increasing, the downward trend shows the series is decreasing, and no trend shows the series is constant. COVID case data was plotted as shown in Figure 2 using Python's statmodels library seasonal decompose to understand the seasonality and noise in the case data and whether the case time series is additive or multiplicative. Seasonality describes the periodic signal in the time series, and the noise signifies the variance and volatility of the time series. The time series plots show that the COVID case trend is non-linear and does not always increase or decrease. Hence, the "multiplicative" trend parameter was used for modeling. The seasonal plot shows an infection surge roughly every six months. Therefore, the seasonality of 6 months (24-week period) was factored into the models.

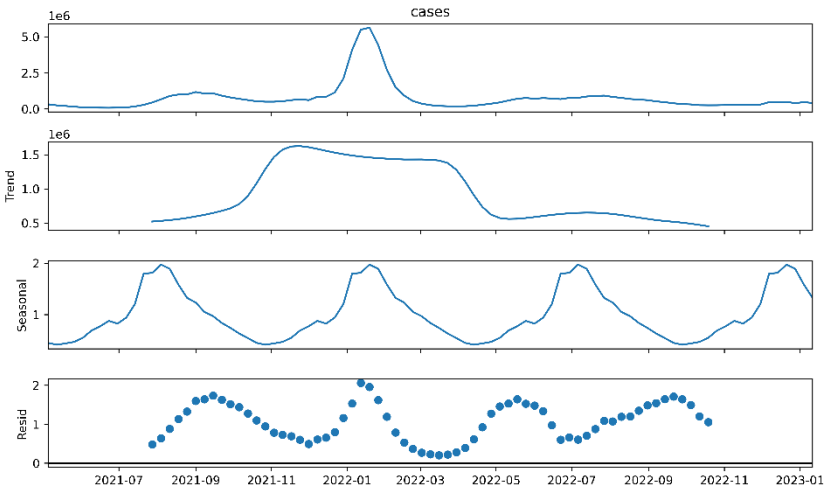

Figure 2. US COVID-19 infection cases. Trend, Seasonality, and Noise.

Models

The model development and coding were performed in Python programming language. Models were implemented using the available libraries and modules: numpy for working with multidimensional arrays, pandas for

working with datasets, matplotlib for data visualization and plotting, statsmodels for conducting statistical modeling and testing, pmdarima for building SARIMAX model, and Sequential, LSTM and Dense modules from keras for building the LSTM model. COVID-19 case data was loaded into a pandas data frame and split into 80% train and 20% test data. Training data was prepared and fitted to the model for prediction. Training data, actual cases, and predicted cases were plotted.

Holt-Winters Model

Holt-Winters exponential smoothing model was first selected for prediction due to its simplicity of implementation. Based on the results from the seasonal decomposition of the COVID-19 case data, the multiplicative trend and seasonality with a seasonal period of 6 months (24 weeks) were considered. An optimized best-fit model fitting approach was used.

SARIMAX Model

The SARIMAX model was chosen next to account for the seasonality of the COVID-19 infection cases. Based on the results from the seasonal decomposition of the COVID-19 case data, the multiplicative trend and seasonality with a seasonal period of 6 months (24 weeks) were considered. Ad-Fuller and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests were run to understand the stationarity. Both tests showed that the COVID-19 infection case data was non-stationary. The case data was transformed using the differencing technique, making it stationary. The model was trained with varying seasonal and non-seasonal orders. The final parameters chosen were (0,0,1) (1,0,1) [24], which had the lowest Akaike information criterion (AIC) values, indicating the best model fit.

LSTM Model

The LSTM model was chosen as it has been reported to be far superior to the ARIMA model in attaining higher prediction accuracy with significantly lower error rates for time-series data prediction (Siami-Namini et al., 2018). The COVID-19 case dataset was normalized using the MinMaxScaler and split into the train and test sets. Train data was then reshaped into a 3D array suitable for input to the LSTM model. Keras Sequential() function was used to build the LSTM model by adding two layers of 50 units each, including a dense output layer. Adam optimizer was used to compile the model. 98 epoch runs with a batch size of 8 were used to train the model. Multiple iterations of the parameter fine-tuning were performed to attain the best prediction results.

Model Performance Evaluation

MAE and RMSE measurements were chosen to evaluate the model's performance due to their applicability in efficiently evaluating errors in time series forecasting. A model with the lowest MAE and RMSE values is the most accurate. MAE measures the average absolute value of the difference between predicted and actual values at a given time. The RMSE measures the average magnitude of the error, where the errors are squared before they are averaged, giving a relatively high weightage to large errors. RMSE is used when large errors are undesirable (Saigal & Mehrotra, 2012).

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |P_t - A_t|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(P_t - A_t)^2}$$

Where A_t is the actual values at the time t, P_t is the predicted values at time t, and n is the number of predictions.

Results

The Holt-Winters model with the exponential smoothing technique was applied using the "multiplicative" trend and seasonal parameters. An optimized set of parameters was used to generate a best-fit model. The training data, actual cases, and model-predicted cases were plotted as shown in Figure 3. The initial few weeks of prediction results looked promising. However, the errors increased significantly towards the end, with the MAE value of 307,444 and the RMSE value of 370,711, both significantly high and showing weak prediction results.

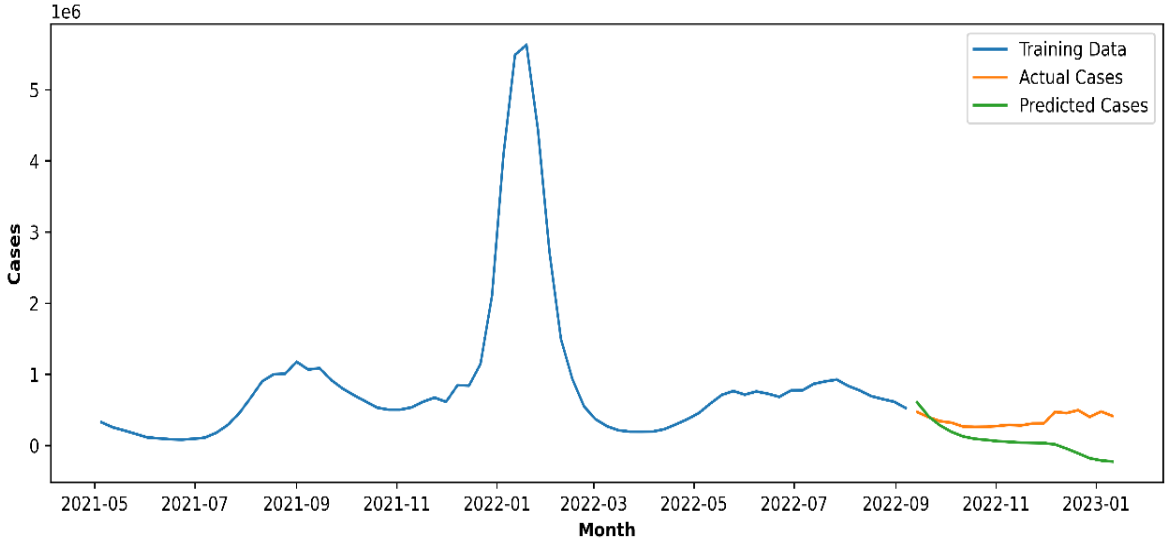

Figure 3. US COVID-19 infection case forecasting using Holt-Winters. Training case data, actual cases, and predicted cases.

SARIMAX model with a "multiplicative" trend, a 24-week seasonal period, and an optimized set of order parameters with the lowest AIC generated the best-fit model. The training data, actual cases, and model-predicted cases were plotted as shown in Figure 4. The SARIMAX model MAE value of 206,052 and RMSE value of 214,947 are lower than the Holt-Winters but still significantly high, showing weak prediction results. The SARIMAX model showed consistently higher prediction accuracy throughout the testing period.

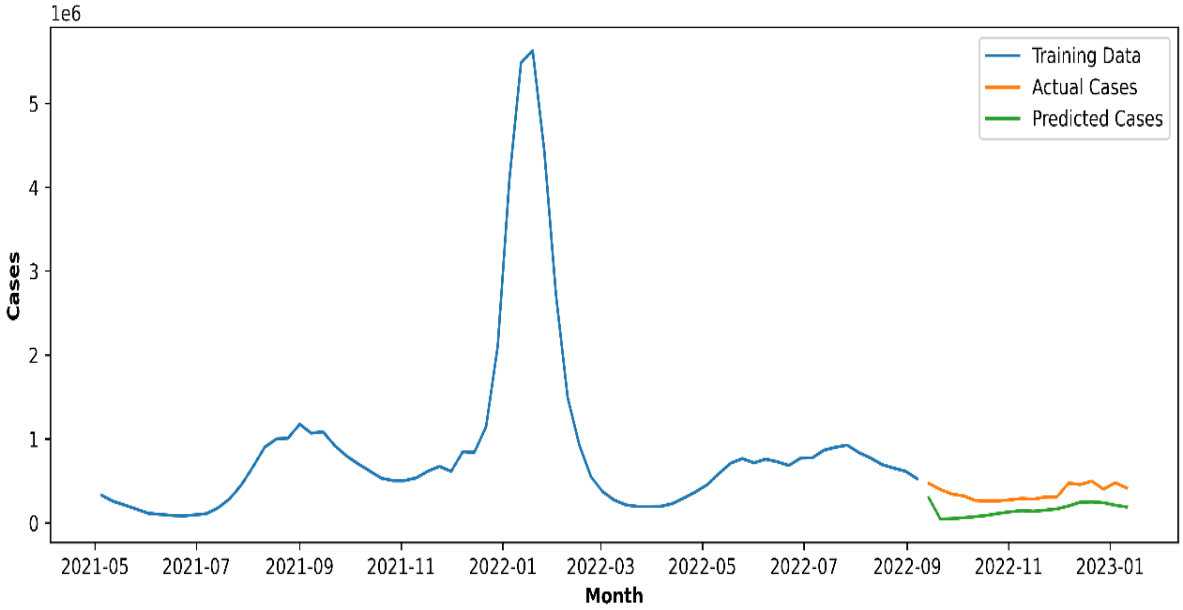

Figure 4. US COVID-19 infection case forecasting using SARIMAX. Training case data, actual cases, and predicted cases.

The LSTM model with Adam optimizer, batch size 8, and 98 epochs provided the best fit. The training data, actual cases, and model-predicted cases were plotted as shown in Figure 5.

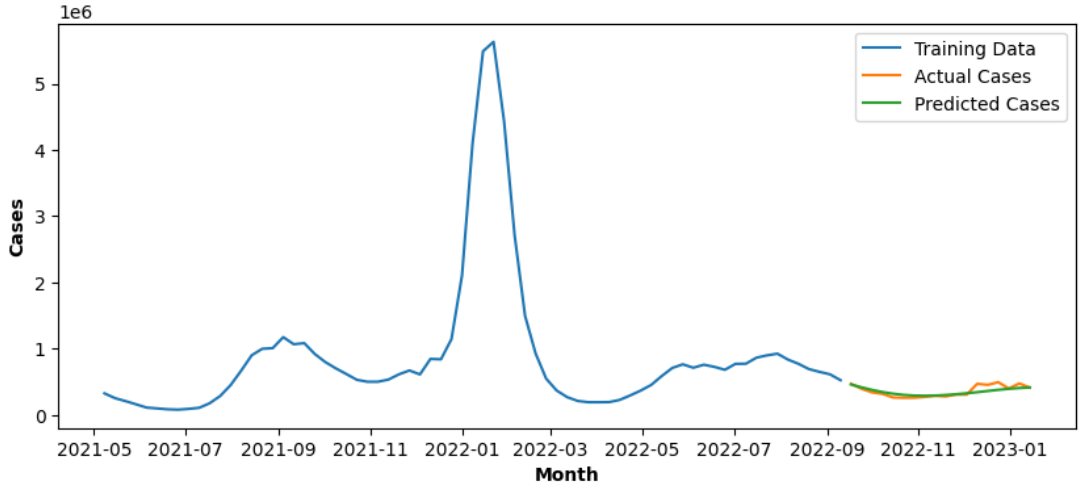

Figure 5. US COVID-19 infection case forecasting using LSTM. Training case data, actual cases, and predicted cases.

The LSTM model had the lowest MAE value of 39,297 and RMSE value of 54,225 compared to the Holt-Winters and SARIMAX models and had the highest prediction accuracy of 91%, as shown in Table 1.

Table 1. COVID Infection Prediction Comparison.

Model	MAE	RMSE	Accuracy
Holt-Winters	307,444	370,711	19%
SARIMAX	206,052	214,947	43%
LSTM	39,297	54,225	91%

Discussion

A significant amount of research has been conducted using multiple methods and modeling techniques since the inception of the pandemic to predict COVID-19 infections. Most of the research framework for COVID-19 predictions has focused on a specific localized set of COVID-19 infection data collected from wastewater epidemiology, hospital x-rays, and historical cases of specific regions or variants using traditional machine learning models. This research presents a very efficient and simplistic COVID-19 infection prediction technique applied to combined variants' infection time-series data using Holt-Winters exponential smoothing, SARIMAX, and LSTM models that have not been explored previously to the author's knowledge. The research was performed on 21 months of US COVID variants aggregated case data. The seasonal plot of COVID cases showed an infection surge roughly every six months attributed to the summer and winter holidays when large people gatherings happen, and more people get infected with the virus, resulting in factoring six months (24-week period) seasonality into the models. The Holt-Winters exponential smoothing model prediction was highly biased towards the last few weeks of train data due to the model's characteristics of the weights decaying exponentially from the most recent to the oldest historical value, which resulted in a degradation of prediction accuracy over time and missing the seasonal impact of COVID variant surges. The SARIMAX model showed higher prediction accuracy, specifically in the seasonality period, than the Holt-Winters model, which is heavily weighted towards the most recent trends. The LSTM model had the best prediction accuracy of 91% and the lowest RMSE value due to its property of selectively remembering patterns for long durations. LSTM has the forget gates that correct the vanishing gradient problem seen in the recurrent neural networks, which minimizes the error losses. This property of LSTM is beneficial for COVID-19 infection modeling, specifically during new emerging variant-related infection surges. It was also noticed that increasing the training times with more epoch runs did not improve the performance of the trained LSTM model. The model can be trained further using more extensive and varied datasets from multiple countries using an optimized set of hyperparameters, including using the bi-directional component of the LSTM to improve accuracy (Sunny et al., 2020). Future work will extend the modeling and prediction benefits to other US regions, countries, and specific hot spots encountering the COVID variant surges.

Conclusion

This study demonstrates the successful deployment of time-series Holt-Winters and SARIMAX machine learning and LSTM RNN deep learning models to forecast combined variant COVID-19 infections with acceptable accuracy. Modeling performed on the US COVID variants aggregated case data showed that the SARIMAX model was more efficient in handling the seasonality in the infection trend than the Holt-Winters model, which is heavily weighted towards the most recent trend. The LSTM model outperformed with the best prediction accuracy of 91% and the lowest RMSE and MAE values due to its property of selectively remembering patterns for long durations. The prediction model can easily be deployed on the COVID-19 cases tracking websites for real-time predictions, enabling swift actions to mitigate the infection surge. This research's findings and the success of the neural network models prompt further research into developing and implementing advanced

deep-learning neural networks in the healthcare industry for accurate, early detection of diseases. Disease modeling and prediction allow for better assessment of different policy decisions and illuminate achievable conditions to mitigate the impact of an epidemic.

Limitations

The LSTM model performance results showcase an effective approach to handling complex datasets, resulting in commendable prediction accuracy. However, the modeling and forecasting were performed on the limited COVID data from the US only, and further expansion to other states, regions, and countries is needed to gain higher confidence and broader acceptance.

Acknowledgments

I sincerely thank my teacher, Kieu Pham at Milpitas High School, for the guidance, support, and opportunity to pursue this research.

References

Alali, Y., Harrou, F., & Sun, Y. (2022). A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models. *Scientific Reports*, *12*(1), 2467. https://doi.org/10.1038/s41598-022-06218-3

Centers for Disease Control and Prevention. (n.d.). *Weekly United States covid-19 cases and deaths by state - archived.* Retrieved November 2023, from https://data.cdc.gov/Case-Surveillance/Weekly-United-States-COVID-19-Cases-and-Deaths-by-/pwn4-m3yp/about_data

Gold, J. A., Rossen, L. M., Ahmad, F. B., Sutton, P., Li, Z., Salvatore, P. P., ... & Jackson, B. R. (2020). Race, ethnicity, and age trends in persons who died from COVID-19—United States, May–August 2020. *Morbidity and Mortality Weekly Report*, *69*(42), 1517. https://doi.org/10.15585%2Fmmwr.mm6942e1

Kannan, S., Shaik Syed Ali, P., & Sheeza, A. (2021). Omicron (B. 1.1. 529)-variant of concern-molecular profile and epidemiology: a mini review. *Eur. Rev. Med. Pharmacol. Sci*, *25*(24), 8019-8022. https://doi.org/10.26355/eurrev_202112_27653

Narin, A., Kaya, C., & Pamuk, Z. (2021). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, *24*, 1207-1220. https://doi.org/10.1007/s10044-021-00984-y

Omane-Adjepong, M., Oduro, F. T., & Oduro, S. D. (2013). Determining the better approach for short-term forecasting of ghana's inflation: Seasonal ARIMA Vs holt-winters. *International Journal of Business, Humanities and Technology*, *3*(1), 69-79.

Polo, D., Quintela-Baluja, M., Corbishley, A., Jones, D. L., Singer, A. C., Graham, D. W., & Romalde, J. L. (2020). Making waves: Wastewater-based epidemiology for COVID-19–approaches and challenges for surveillance and prediction. *Water research*, *186*, 116404. https://doi.org/10.1016/j.watres.2020.116404

Saigal, S., & Mehrotra, D. (2012). Performance comparison of time series data using predictive data mining techniques. *Advances in Information Mining*, *4*(1), 57-66.

Shrestha, L. B., Foster, C., Rawlinson, W., Tedla, N., & Bull, R. A. (2022). Evolution of the SARS-CoV-2 omicron variants BA. 1 to BA. 5: implications for immune escape and transmission. *Reviews in Medical Virology*, *32*(5), e2381. https://doi.org/10.1002/rmv.2381

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019, December). The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International conference on big data (Big Data)* (pp. 3285-3292). IEEE. https://doi.org/10.1109/BigData47090.2019.9005997

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018, December). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1394-1401). IEEE. https://doi.org/10.1109/ICMLA.2018.00227

Sunny, M. A. I., Maswood, M. M. S., & Alharbi, A. G. (2020, October). Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. In *2020, 2nd novel intelligent and leading emerging sciences conference (NILES)* (pp. 87-92). IEEE. https://doi.org/10.1109/NILES50944.2020.9257950

World Health Organization (WHO) *Coronavirus (COVID-19) dashboard*. Retrieved November 2023, from https://covid19.who.int/

Wortham, J. M., Lee, J. T., Althomsons, S., Latash, J., Davidson, A., Guerra, K., ... & Reagan-Steiner, S. (2020). Characteristics of Persons Who Died with COVID-19-United States, February 12-May 18, 2020. *MMWR. Morbidity and mortality weekly report*, *69*(28), 923-929. http://dx.doi.org/10.15585/mmwr.mm6928e1

Zou, D., Wang, L., Xu, P., Chen, J., Zhang, W., & Gu, Q. (2020). Epidemic model guided machine learning for COVID-19 forecasts in the United States. *MedRxiv*, 2020-05. https://doi.org/10.1101/2020.05.24.20111989