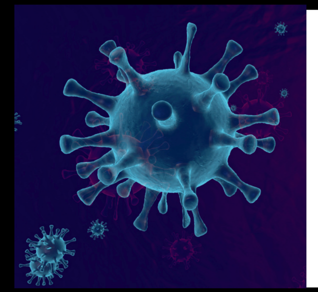


Bridging the Gap: Enhancing COVID-19 Epidemic Forecasting by Integrating Factors like Vaccination Rates, Mobility, Stringency, Socio-Economic Indicators, and Health Metrics into Time Series Models

Authors

Akshat Karwa
akshatkarwa@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Mehul Rastogi
mehulrastogi@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Main Idea

Pandemic spread is a multi-dimensional phenomenon.

There is a need for models to be context-aware and adaptable.

Important to understand and integrate social, economic, and behavioural factors to improve forecasting and gain insights.

Dataset Details

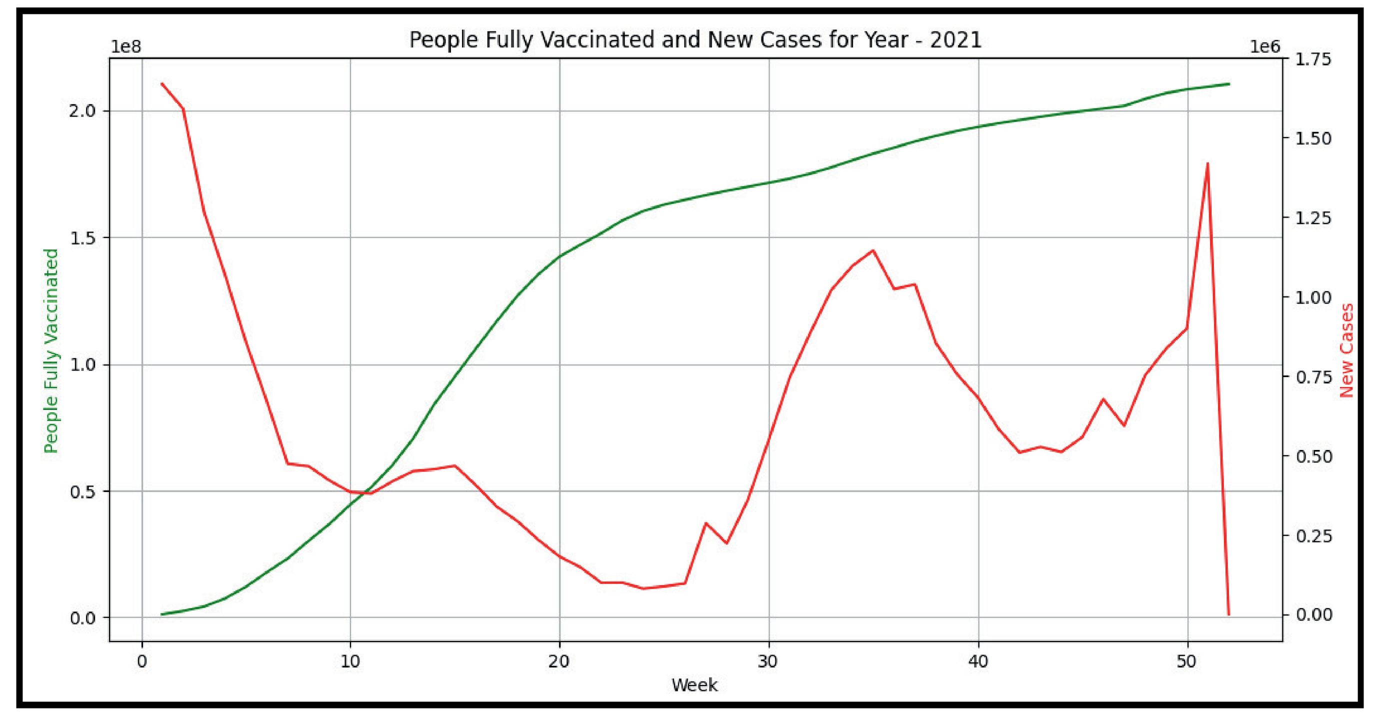
Two primary datasets were utilised:

- **WHO COVID-19 Dashboard Data** for the United States (1,674 days, 68 features)
- **COVID-19 Twitter Social Mobility Data**

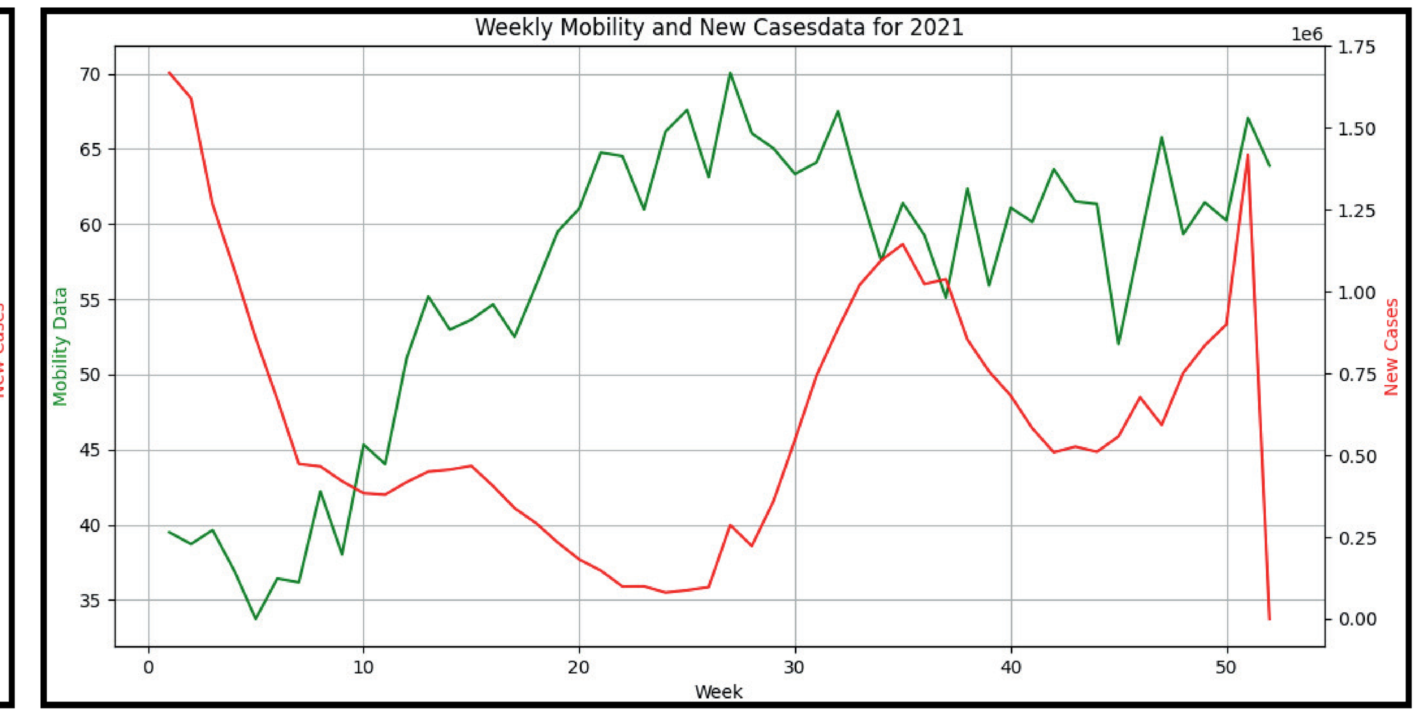
Post merging & feature reduction, final dataset spanned 1,097 days from January 5, 2020, to December 31, 2022, with 38 features.

A **70-30 train-test data split** was used for model learning and performance assessment across the various models.

Data Insights



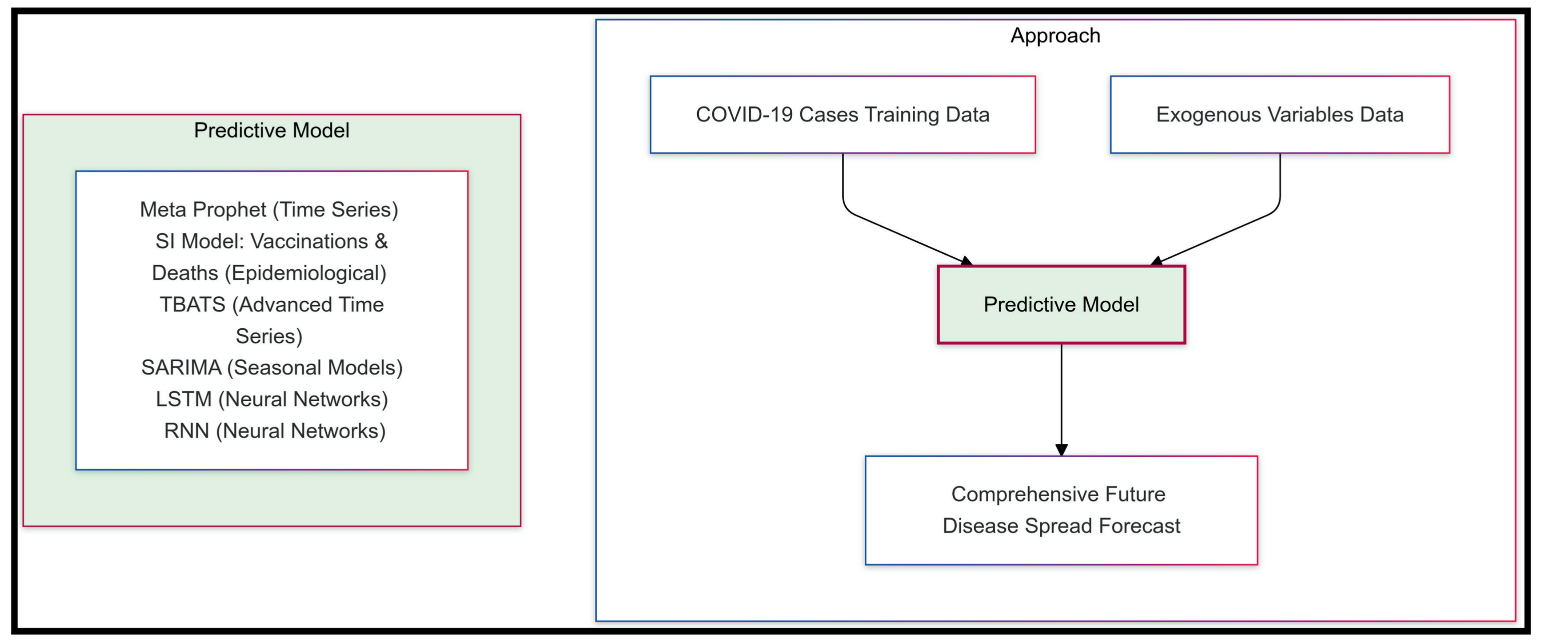
COVID New Cases vs Vaccination Trends

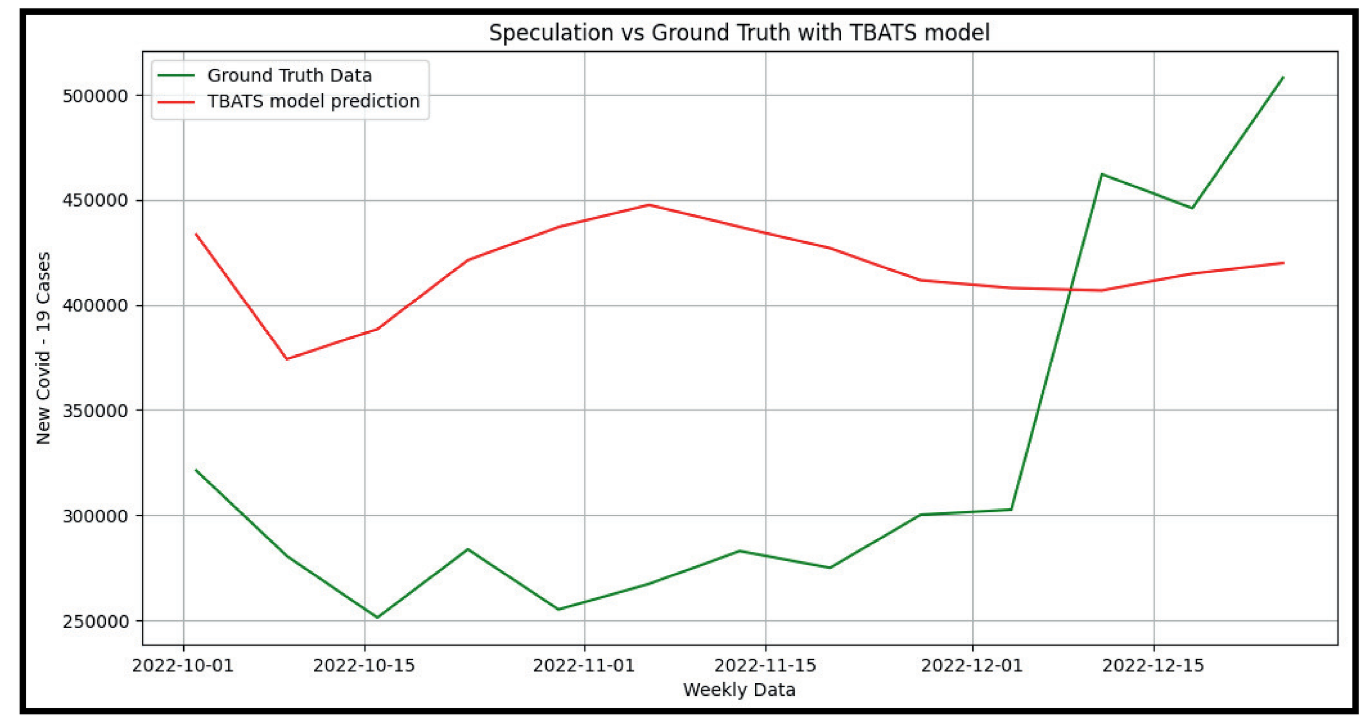


COVID New Cases vs Social Mobility Trends

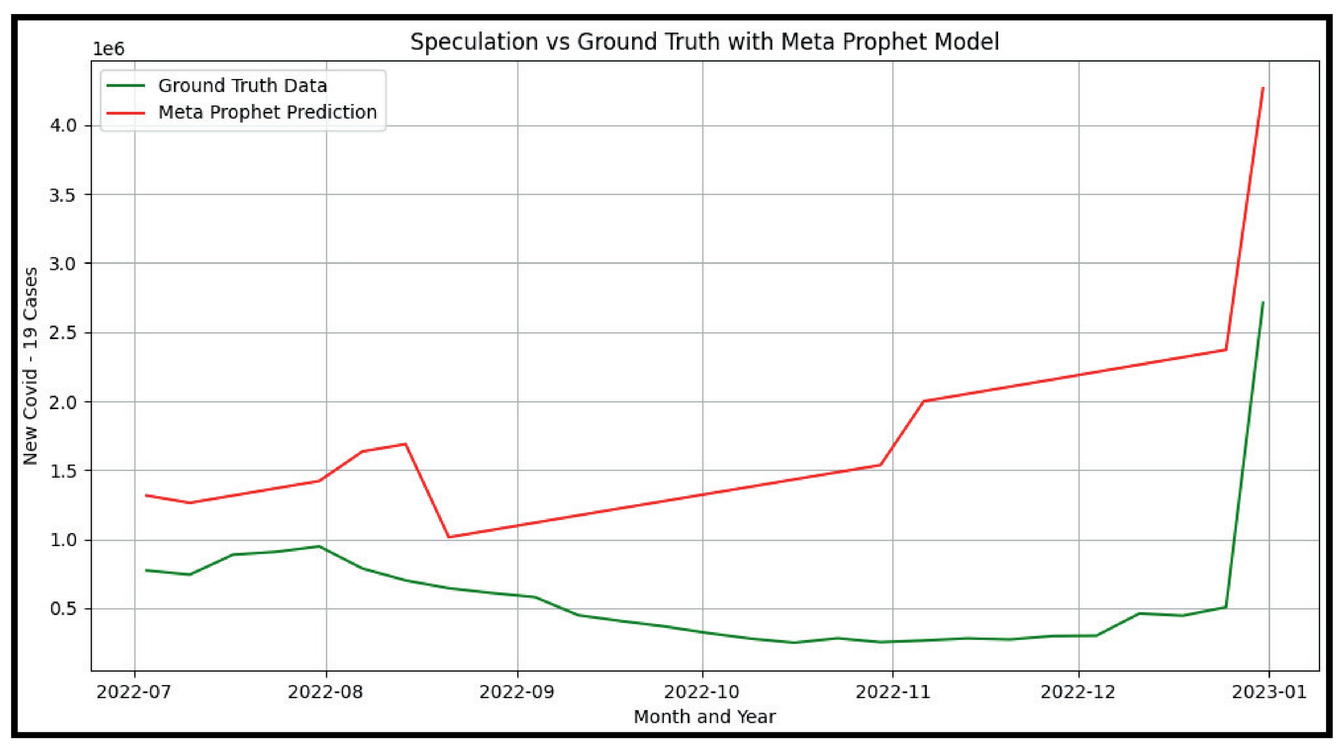
Exogenous Variables

Human Development Index (HDI): Socioeconomic resilience indicator.
GDP Per Capita: Economic healthcare capacity
Diabetes Prevalence: Population health vulnerability
Vaccinations: Immunisation progress tracker
Social Mobility Index: Population movement patterns
Stringency Index: Containment strategy effectiveness





- **TBATS (Trigonometric, Box-Cox, ARMA, Trend, and Seasonal)** Model excellently captures complex seasonal patterns in the time series data.
- Offers accurate weekly predictions and interpolates missing data to provide precise forecasts.
- Comprehends seasonal variations better than SARIMAX and even without incorporating exogenous variables it performs really accurately.



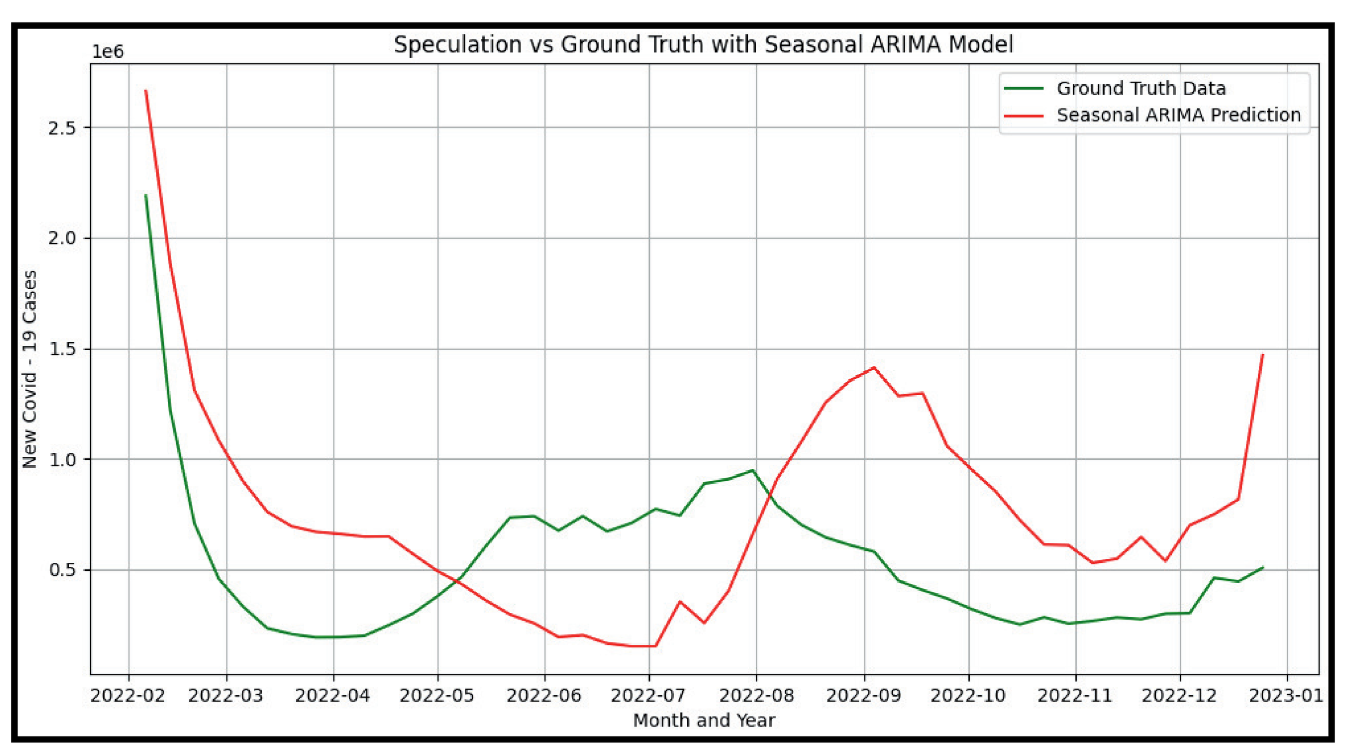
Facebook Prophet Model - Initialised with weekly seasonality.

Exogenous Variable Combinations: a) HDI, Mobility **b)** Vaccinations, Stringency **c)** Stringency, GDP Per Capita **d)** Mobility, Vaccinations, Stringency, HDI, Diabetes, GDP Per Capita

Best Performing Model: (b)

Performance Insights: Strong correlation between vaccination and cases. Better performance with fewer exogenous variables. Excellent at depicting rises and falls in cases.

GDP Per Capita > Diabetes Prevalence > HDI > Vaccinations > Stringency > Mobility



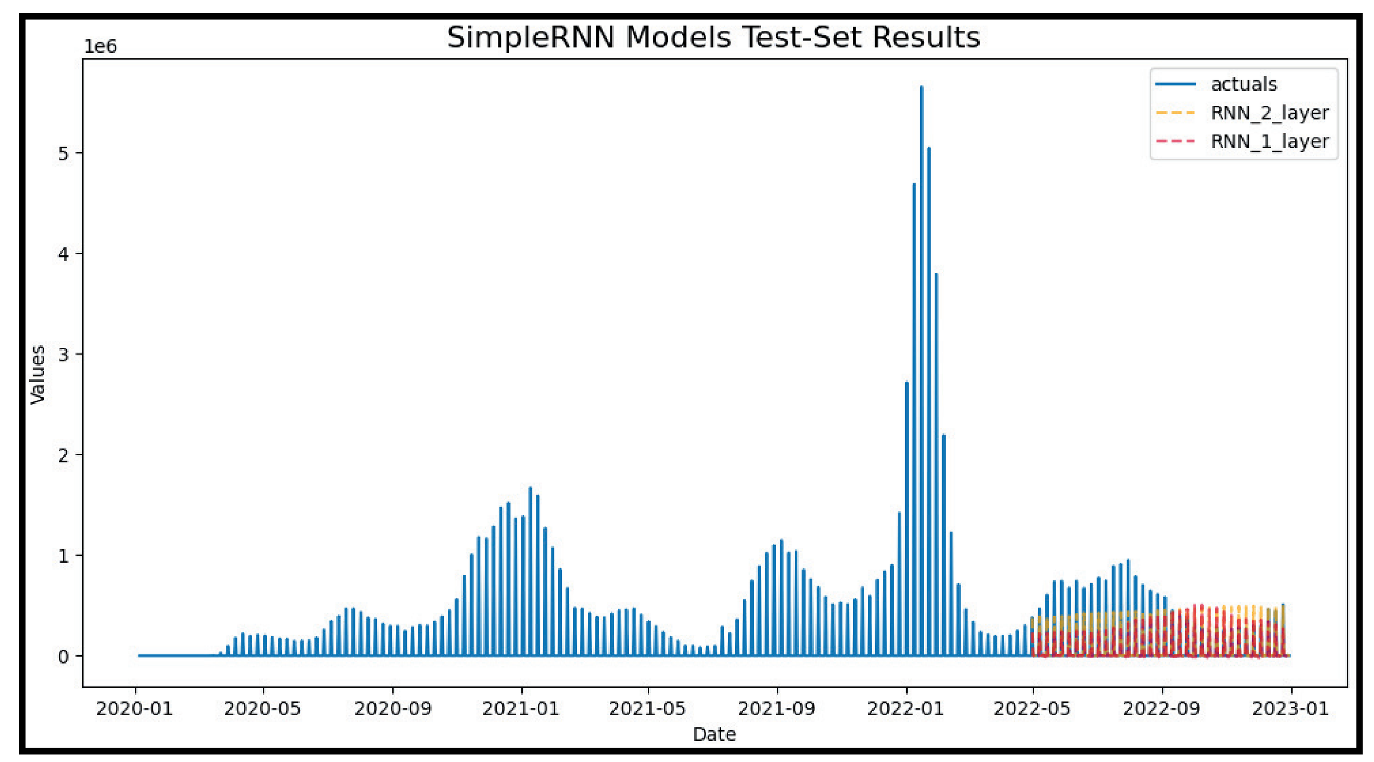
Seasonal ARIMA (Order (1,1,1) & Seasonal order (1,1,1,52))

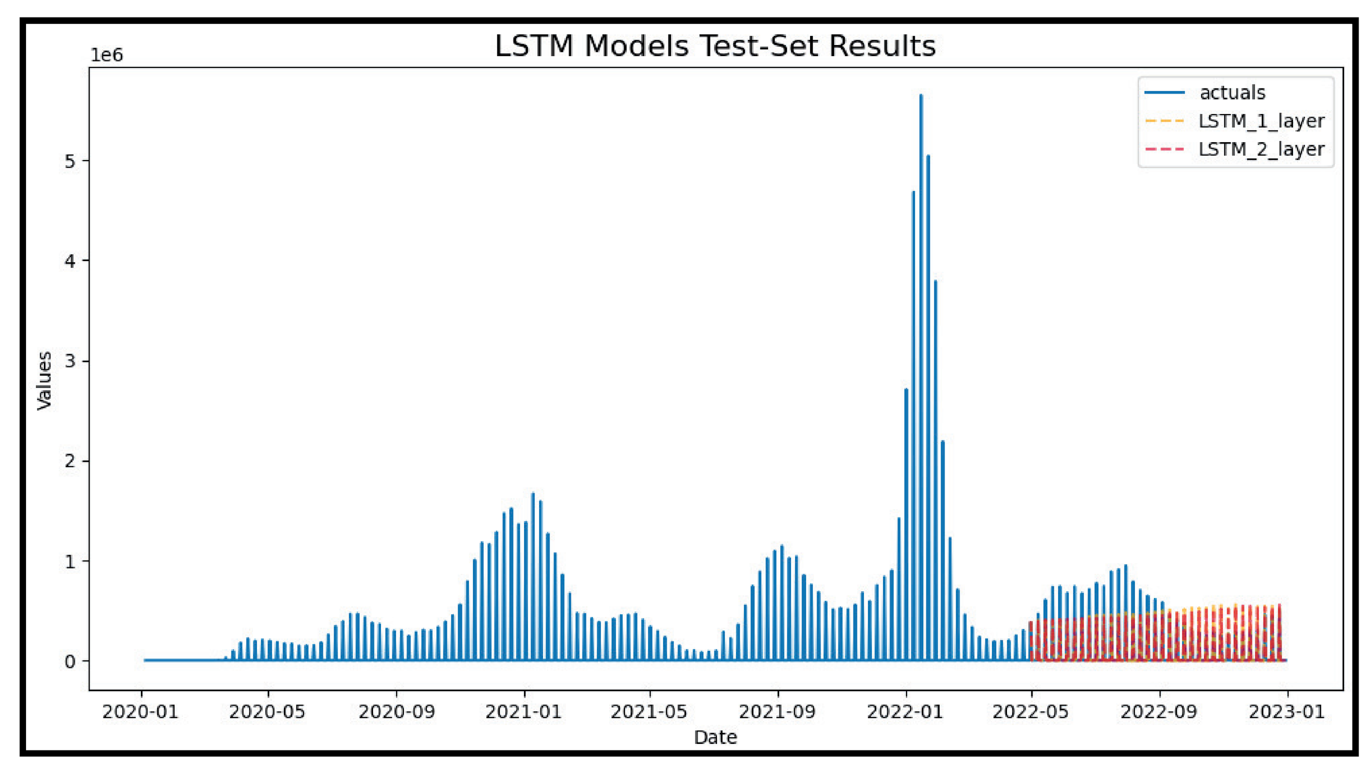
Exogenous Variable Combinations: (a) Mobility **(b)** Mobility, Stringency **(c)** Mobility, Stringency, GDP Per Capita **(d)** Mobility, GDP Per Capita, Vaccinations

Best Performing Model: (d)

Performance Insights: Mobility is more important than Stringency (somewhat linearly dependent). GDP Per Capita in (c) slightly worsened the prediction accuracy. In (d), many exogenous variables caused hallucinations.

Vaccinations > Human Development Index > Diabetes Prevalence > GDP Per Capita > Mobility Data > Stringency Index





- **LSTM** and **SimpleRNN** effectively capture overall trends.
- Follow the broader trajectory.
- Not able to predict sharp peaks.
- RNN-based models were able to learn temporal dependencies, and understand underlying patterns.
- Integrating external factors into RNN models can provide further insights and enhance forecasting.
- The **RNN_2_Layer** model outperformed others. The LSTM_1_Layer model achieved strong results.
- LSTM_2_Layer showed higher error rates, emphasizing the need for more balanced architecture.

Model Name	Test Set RMSE	Test Set MAE
RNN_2_layer	95,198.32	36,336.57
LSTM_1_layer	98,318.51	34,102.96
RNN_1_layer	133,052.04	47,039.23
LSTM_2_layer	216,635.11	77,186.12