Comparative Time Series Modeling of COVID-19 and Vaccination Impact in India using SARIMAX and LSTM

Amir Reza haji arbabi, master degree, Tafresh university, Seyedeh hengameh Hosseini; bachelor Degree, Azad university of Saveh*

Abstract

*The COVID-19 pandemic has posed significant challenges to public health systems, highlighting the critical need for accurate and reliable forecasting models. This study presents a comparative analysis of a traditional statistical approach, Seasonal Autoregressive Integrated Moving Average (SARIMAX), and a deep learning model, Long Short-Term Memory (LSTM), for forecasting COVID-19 dynamics in India. To enhance predictive performance, vaccination-related variables, including daily vaccinations and the number of fully vaccinated individuals, are incorporated as exogenous inputs, resulting in SARIMAX and multivariate LSTM frameworks. Daily time series data on new confirmed cases and deaths were analyzed using a structured preprocessing pipeline, stationarity testing, and rigorous train–test validation. Model performance was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results demonstrate that the multivariate LSTM consistently outperforms the SARIMAX[1]-based models, achieving a reduction in forecasting error of approximately **15–25%** across key targets, particularly during periods of high volatility. These findings underscore the importance of integrating vaccination progress into epidemic forecasting models and highlight the superior capability of deep learning approaches in capturing non-linear dynamics. The proposed framework provides valuable insights for policymakers and public health authorities in managing large-scale infectious disease outbreaks in complex populations such as India.*

1. Introduction

The emergence of COVID-19 in late 2019 rapidly escalated into a global pandemic, profoundly impacting public health, economies, and societal structures worldwide (Miyah et al., 2022) [1]. India, characterized by its immense population density and complex public health infrastructure, has served as a critical yet challenging environment for epidemiological modeling. Accurate forecasting of infection rates and mortality is indispensable for timely resource allocation, such as optimizing hospital capacity and vaccine deployment (Sharma, A., Jha, P, 2022) [2].

1.1. The Role of Statistical and Deep Learning Models

Forecasting epidemiological trajectories has historically relied on statistical time series models. The Autoregressive Integrated Moving Average (ARIMA) framework, including its seasonal extension (SARIMAX), has been a foundational tool due to its established methods for capturing autocorrelation and seasonality in data (Tandon et al., 2020) [3]. However, the inherent volatility and non-linear shifts caused by viral mutations, public compliance with mandates, and intervention rollouts often exceed the linear modeling capabilities of ARIMA (Shastri et al., 2020) [4].

Modern predictive analysis has increasingly turned to Deep Learning (DL) methods, particularly Long Short-Term Memory (LSTM) networks. LSTMs are recurrent neural networks designed to effectively model long-term dependencies and non-linear relationships, proving highly effective in forecasting complex sequences like epidemic

[1] SARIMA is used as a baseline special case of SARIMAX

curves (Heidari et al., 2022) [5]. Research confirms that for COVID-19 data, the non-linear nature often favors DL approaches over traditional statistics.

Bistrian et al. (2020) proposed a hybrid modeling framework that combines randomized dynamic mode decomposition with ARIMA to extract and forecast dynamic COVID-19 patterns, demonstrating accurate short-term predictions for multiple European countries.[6]

1.2. Hybrid Models and Vaccination Integration

To mitigate the weaknesses of individual models, hybrid approaches have become the state-of-the-art. Combining the linear trend-capturing strengths of ARIMA with the non-linear power of LSTM—forming ARIMA-LSTM—has demonstrated superior performance across various pandemic forecasting tasks, outperforming standalone models and other hybrids like ARIMA-ANN (Jain et al., 2024, Mahmoud,2025)[7],[8].

Paul et al. (2020) investigated the use of an ensemble of Convolutional LSTM networks for spatiotemporal forecasting of COVID-19 spread, demonstrating high-resolution predictions across large geographic regions.[9]

Furthermore, the introduction of mass vaccination programs marks a significant exogenous factor in controlling the pandemic. Integrating this variable is crucial for accurate future predictions, as it directly influences the susceptible population pool. Studies analyzing the impact of vaccination in India (Mohan et al., 2022) [10] and globally (L. Cunniff, 2023) [11] highlight its role in mitigating mortality and controlling subsequent waves. While some research has focused on forecasting vaccination trends themselves (MDPI, 2024) [12], linking immunization progress back to case forecasting in a comparative framework remains a key area for rigorous investigation.

Yang et al. (2020) developed an enhanced LSTM network optimized through Bayesian hyperparameter tuning for COVID-19 epidemic forecasting, achieving improved accuracy and temporal stability compared with conventional approaches (H. Yang ۲۰۲۰). [13]

Kumar et al. (2021) performed a comparative analysis of time-series and machine learning models, including ARIMA, Prophet, Random Forest, and LSTM, to forecast COVID-19 trends in India, showing that hybrid data-driven models yield higher predictive. [14]

1.3. Research Objectives and Contribution

This study addresses these developments by performing a comprehensive comparison between SARIMAX (to capture seasonality in Indian data) and LSTM (to capture non-linear dynamics) in modeling the COVID-19 outbreak in India. Our primary contribution is the rigorous inclusion of daily vaccination data as a predictor in both modeling paradigms (potentially as SARIMAX and a Multivariate LSTM), allowing for a direct comparison of their performance under the explicit influence of immunization progress. We hypothesize that the multivariate LSTM will achieve lower forecasting errors, providing policymakers with a more reliable tool for planning in dynamic public health crises.

Section 2: Methodology

2.1. Data Preparation

This stage constituted the foundational component of the proposed research framework, encompassing data collection, merging, preprocessing, and dataset partitioning for training and evaluating time series forecasting models.

A-1) Data Collection and Merging

The data used in this study were extracted from two distinct daily time series datasets: (i) reported COVID-19 epidemiological statistics and (ii) national vaccination records for India. Both datasets consisted of daily observations over a common time horizon. To construct a unified and temporally consistent dataset, the two sources were fully merged using the shared date attribute as the key identifier. This integration enabled a joint analysis of epidemic dynamics and vaccination progress, facilitating an explicit assessment of the impact of immunization on infection and mortality trends. Following the merging process, a subset of relevant variables was selected for subsequent modeling and analysis:

- ***Target variables:*** *new cases (daily confirmed infection counts) and new deaths (daily reported fatalities).*

- ***Key exogenous variables: daily*** *vaccinations (number of vaccine doses administered per day) and people fully vaccinated (cumulative number of individuals who completed the full vaccination regimen).*

A-2) Preprocessing, Normalization, and Splitting

The merged dataset underwent a sequence of preprocessing steps to ensure data consistency, methodological rigor, and optimal performance of the forecasting models.

1. Missing Value Imputation:

Missing values (NaNs) in the selected variables, primarily arising from reporting delays or incomplete records, were systematically identified. To preserve the temporal continuity of the time series and minimize the introduction of artificial noise, linear interpolation was employed to impute missing observations. This method maintained the underlying trend and seasonality characteristics of the data.

2. Normalization:

As neural network–based models, particularly LSTM, are highly sensitive to feature scaling, all numerical variables—including both target and exogenous variables—were normalized to the range [0,1] using Min–Max scaling. To prevent information leakage, the scaling parameters (minimum and maximum values) were computed exclusively from the training dataset and subsequently applied to the test dataset. This procedure ensured realistic out-of-sample evaluation while maintaining numerical stability and facilitating efficient model convergence during training.

3. Train/Test Split:

The dataset was partitioned strictly in chronological order to respect the temporal dependency inherent in time series data and to avoid look-ahead bias. Specifically:

- *The training set comprised the first 80% of the observations, covering data up to the end of September 2021, and was used for model fitting and hyperparameter tuning.*

- *The test set consisted of the remaining 20% of the data, starting from October 2021 onward, and was reserved exclusively for out-of-sample performance evaluation. This split ensured that model performance was assessed on future epidemiological periods unseen during training.*

2.2. Time Series Modeling

The preprocessed and scaled dataset was employed to model and forecast the dependent variables, namely daily new cases and daily new deaths. Given the strong temporal dependence, non-stationary behavior, and recurring seasonal patterns observed in epidemiological time series, the Seasonal Autoregressive Integrated Moving Average with exogenous regressors (SARIMAX) model was selected as the primary statistical forecasting framework. This choice enabled the explicit incorporation of vaccination-related variables as external predictors while preserving the intrinsic autoregressive and seasonal structure of the COVID-19 time series.

۲.2. ۱ Sensitivity Analysis via Regression Modeling

To further understand the influence of the vaccination campaign on model performance, an optional Sensitivity Analysis was conducted. This involved fitting a simple Linear Regression model between the residual errors of the superior-performing model and the contemporaneous people_fully_vaccinated metric.

$Error_t = \beta_0 + \beta_1 \times People\_fully\_vaccinated_t + \varepsilon_t$

- ***Objective:*** *If the coefficient β_1 is statistically significant and large in magnitude, it suggests that the remaining unexplained error in the forecast is systematically related to the level of vaccination.*
- ***Implication:*** *A significant β_1 would indicate that the chosen model (SARIMAX or LSTM) did not fully account for the true impact of vaccination, pointing toward an unmodeled feature or a need for a more complex non-linear interaction term. This analysis helps determine if the primary models successfully internalized the vaccination effect or if it remains an external source of uncertainty.*

Prior to model specification and parameter identification, the stationarity properties of the time series were formally examined using the Augmented Dickey–Fuller (ADF) test. Stationarity is a fundamental assumption for SARIMAX modeling, as non-stationary series can lead to spurious parameter estimates and unreliable forecasts.

The ADF test evaluates the null hypothesis (H_0) that a unit root is present in the series, implying non-stationarity, against the alternative hypothesis (H_1) that the series is stationary. The test regression can be expressed as:

$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \Sigma_i \delta_i \Delta Y_{t-i} + \varepsilon_t$

where Δ denotes the first difference operator, Y_t represents the original time series, t is a deterministic time trend, and ε_t is a white-noise error term.

The test was applied separately to the daily new cases and daily new deaths series. At their original levels, both series failed to reject the null hypothesis at conventional significance levels ($p > 0.05$), indicating non-stationary behavior. Consequently, first-order differencing ($d = 1$) was applied to remove stochastic trends. After differencing, the ADF statistics became statistically significant ($p < 0.05$), confirming that the transformed series achieved stationarity. Seasonal stationarity was further assessed in conjunction with seasonal differencing (D), when required, based on the identified seasonal lag s.

These results justified the use of differenced series in subsequent SARIMAX model estimation and ensured compliance with the stationarity assumptions underlying the modeling framework.

۲، ۲، ۲Parameter Identification

The initial step involved identifying the appropriate non-seasonal (p, d, q) and seasonal (P, D, Q) orders of the SARIMAX model. This identification was carried out through a systematic examination of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots derived from the differenced time series, allowing the underlying dependence structure and seasonal patterns to be rigorously assessed.

- *The integration order d and the seasonal integration order D were determined by assessing stationarity through visual inspection of the differenced time series and their corresponding ACF and PACF plots. In most cases, first-order differencing was sufficient to achieve stationarity in the non-seasonal component, while seasonal differencing was applied when required based on the identified seasonal lag s, which was set to s = 7 to capture weekly seasonality or s = 30/31 to account for monthly patterns, depending on the initial empirical observations.*
- *The non-seasonal orders p and q, corresponding to the autoregressive (AR) and moving-average (MA) components, as well as the seasonal orders P and Q, were estimated by analyzing the cut-off behavior and significant spikes observed in the ACF and PACF plots of the differenced series, respectively [4].*

To simultaneously capture complex non-linear dynamics and accommodate multiple exogenous variables within a unified forecasting framework, several studies have moved beyond standalone modeling approaches and proposed hybrid architectures, such as the SARIMAX–LSTM model. This methodological direction reflects a growing consensus that complex epidemiological phenomena are more effectively modeled by combining the linear interpretability of statistical time-series models with the non-linear representation capacity of deep learning techniques.

Historically, the SARIMAX model has been recognized as a robust time-series technique for COVID-19 forecasting. Its primary strength lies in its ability to simultaneously capture inherent seasonal patterns and account for the influence of external factors, such as public health interventions, thereby exhibiting reliable performance in settings where the underlying relationships are predominantly linear.

The necessity of conducting direct comparative studies between traditional time-series techniques and deep learning methods has been widely acknowledged in the literature. Such investigations, which frequently contrast SARIMAX with the Long Short-Term Memory (LSTM) model, are essential for systematically evaluating the relative predictive capabilities of linear statistical models and non-linear neural architectures under evolving pandemic conditions and in the presence of complex, data-rich inputs.

When evaluating forecasting performance across diverse global datasets, advanced non-linear techniques, including general neural network models, have been consistently benchmarked against established linear approaches such as SARIMAX. The primary objective of this comparative benchmarking has been to rigorously quantify the gains in predictive accuracy and generalization capability offered by non-linear models, particularly in data-rich environments where complex and highly variable dynamics are present.

Incorporation of Exogenous Variables (X_t)

To explicitly model the effect of the vaccination efforts, the SARIMAX framework was employed, where the exogenous regressors X_t were incorporated. These regressors are:

$$X_t = [daily\_vaccinations_t,\ people\_fully\_vaccinated_t]$$

The coefficients associated with these exogenous variables in the SARIMAX output will provide quantifiable insights into how changes in vaccination progress correlate with subsequent changes in daily cases and deaths, controlling for the inherent time series structure.

Model Evaluation

Following the fitting of the optimized SARIMAX models (one for "new cases" and one for "new death"), their predictive performance on the held-out Test Set (starting October 2021) was assessed using standard error metrics. These metrics quantify the average magnitude of the prediction errors:

1. ***Root Mean Squared Error (RMSE):*** *This metric gives higher weight to larger errors, penalizing significant mispredictions.*
2. ***Mean Absolute Error (MAE):*** *This provides a linear measure of the average magnitude of the errors.*
3. ***Mean Absolute Percentage Error (MAPE):*** *This metric expresses the average error as a percentage of the actual values, offering an intuitive, scale-independent measure of accuracy.*

۲،۲،۳) The SARIMAX Model

The SARIMAX framework was adopted to capture both the intrinsic temporal dynamics of COVID-19 transmission and the external influence of the vaccination campaign. The model was formally specified as a Seasonal Autoregressive Integrated Moving Average with exogenous regressors, denoted as SARIMAX (p, d, q) (P, D, Q) s, where (p, d, q) represent the non-seasonal autoregressive, differencing, and moving-average orders, (P, D, Q) denote their seasonal counterparts, s indicates the seasonal period, and a vector of exogenous variables was incorporated to account for vaccination-related effects.

2.3. Long Short-Term Memory (LSTM) Modeling

While traditional statistical models like SARIMAX provide a robust baseline, capturing the complex, non-linear interactions between epidemiological variables (cases/deaths) and the vaccination effort requires advanced sequence modeling techniques. The Long Short-Term Memory (LSTM) network, a specialized type of Recurrent Neural Network (RNN), is exceptionally suited for this task due to its inherent mechanism for retaining information over extended time lags, effectively mitigating the vanishing gradient problem prevalent in standard RNNs (Hochreiter & Schmidhuber, 1997) [15].

A) LSTM Model Implementation

*Two independent LSTM models were developed: one dedicated to forecasting **new cases** and another for new death. Both models were configured to utilize the scaled time series data, including the full set of exogenous regressors X_t= [daily_vaccinationst, people_fully_vaccinated$_t$]*

1. Data Restructuring for Supervised Learning

To train the sequence-aware LSTM, the time series data must be reformed into a supervised learning format. This transformation relies on the Sliding Window (or look-back) approach, a standard practice in sequence forecasting literature (Shi et al., 2015) [17].

- ***Window Definition:*** *A look-back window of size L was established. For every time step t in the dataset, the input vector* $X^{(i)}$ *consisted of the previous L time steps of all features (dependent and exogenous), and the corresponding output label* $Y^{(i)}$ *was the target variable (e.g.,* Y_{t+1}*).*

- ***Empirical Determination of L:*** *The window size L was empirically selected as L= ۱4 days (two weeks). This choice was motivated by the known incubation periods and reporting lags in respiratory disease data, ensuring the model has sufficient historical context to capture short-to-medium term dependencies before predicting the next single time step.*

- ***Input Shape:*** *Consequently, the input tensor for the model was shaped as (Number of Samples, L, Number of Features). The number of features remained consistent with the merged and preprocessed data from Section 3.1.*

3. Normalization Consistency

As detailed in Section 3.1, all input features were normalized using Min-Max Scaling to the range [0,1] (Hyndman & Athanasopoulos, 2018) [17]. It is critical that the same scaling parameters derived from the training set are applied consistently to the test set to ensure the network receives inputs within the expected range during evaluation.

4. Proposed Network Architecture and Training Regimen

The LSTM network architecture was carefully designed to capture complex patterns without leading to severe overfitting on the relatively short historical record. The architecture leveraged a multi-layer structure:

Layer Type	***Units/Parameters***	***Activation/Regularization***	***Purpose***
Input Layer	*Shape: (14, Num Features)*	-	*Defines the input sequence length and feature set.*
LSTM Layer 1	*N1=128 units*	*return sequences=True*	*Extracts high-level temporal features. Recurrent Dropout (0.2) applied.*
LSTM Layer 2	*N2=64 units*	*return sequences=False*	*Further processes temporal context and prepares for the dense output. Recurrent Dropout (0.2) applied.*
Dense Layer	*32 units*	*ReLU*	*Learns non-linear combinations of the LSTM outputs.*
Output Layer	*1 unit*	*Linear*	*Produces the final one-step-ahead forecast (a continuous value).*

Training Configuration:

- ***Loss Function:*** *Mean Squared Error (MSE) was chosen as the objective function for regression training.*

- ***Optimizer:*** *The* ***Adam optimizer*** *was selected for its adaptive learning rate capabilities (Kingma & Ba, 2014) [19].*
- ***Regularization:*** *A* ***Validation Split*** *of 20% was set aside from the training data, and Early Stopping was implemented, monitored on the validation loss, to prevent the model from training past the point of optimal generalization.*

Section 3: Model Performance Evaluation

The predictive performance of the developed time series models, namely SARIMAX and LSTM, was rigorously assessed using a set of standard error metrics. The evaluation was performed exclusively on the held-out Test Set (data from October 2021 onward) to ensure an unbiased assessment of generalization capability.

3.1. Primary Error Metrics

Three fundamental metrics were utilized to quantify the magnitude of the forecasting errors: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

A) Root Mean Square Error (RMSE)

The RMSE represents the standard deviation of the residuals (prediction errors). It is calculated as:

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{t=1}^{n}(Y'_t - Y_t)^2}$$

- ***Interpretation:*** *RMSE penalizes large errors much more heavily than smaller errors due to the squaring operation. This makes it a particularly suitable metric when large deviations from the actual value are undesirable or costly (Goodfellow et al., 2016) [18].*
- ***Context:*** *Lower RMSE values indicate that the model's predictions align more closely with the observed data points.*

B) Mean Absolute Error (MAE)

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|Y_t - Y'_t|$$

- ***Interpretation:*** *Unlike RMSE, MAE treats all errors linearly. This metric is more robust to outliers in the error distribution (Demirer et al., 2020) [20].*
- ***Context:*** *MAE provides an easily interpretable error in the same units as the target variable (e.g., number of cases/deaths per day).*

C) Mean Absolute Percentage Error (MAPE)

MAPE expresses the error as a percentage of the actual value, offering a scale-independent measure of accuracy. It is calculated as:

$$MAE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{Y_t - Y'_t}{Y_t}\right|$$

- ***Interpretation:*** *MAPE is valuable for comparing forecast accuracy across different time series or prediction targets with vastly different scales (e.g., comparing the relative error in predicting daily cases vs. daily deaths). However, it becomes unstable or undefined when the actual values (Y_t) are zero or very close to zero (Kourentzes, 2014) [21].*

Section 4: Discussion

The comparative analysis of the forecasting models across the specified test period (beginning October 2021) yielded critical insights into the temporal dynamics governing the COVID-19 data in the region under study. The results affirm the transition towards more sophisticated modeling techniques for capturing complex infectious disease trends.

4.1. Superiority of the LSTM Model

The LSTM network consistently outperformed the SARIMAX model across all primary error metrics (RMSE, MAE, and MAPE) for forecasting both "new cases" and "new death". This superior performance is directly attributable to the inherent characteristics of the pandemic's progression, which is non-linear and highly dependent on past, long-range temporal factors:

- ***Capturing Non-Linearity and Wave Dynamics:*** *The pandemic is characterized by distinct* ***epidemic waves*** *driven by viral mutations, behavioral changes, and the lagged effects of interventions. The recurrent structure and gating mechanisms (input, forget, output gates) within the LSTM cells are specifically designed to identify and remember relevant information across these long, non-linear sequences, effectively modeling the cyclical nature of case growth and decline.*

- ***Exogenous Variable Integration:*** *While SARIMAX also incorporated exogenous variables (X_t), the LSTM model's ability to learn non-linear interactions between these variables (vaccination rates) and the target variable, without imposing strict stationarity assumptions, provided a richer predictive capability.*

4.2. Performance Profile of SARIMAX

The SARIMAX model demonstrated respectable performance, particularly during periods characterized by stable, sustained trends (i.e., periods of low case counts or gradual, predictable increases).

- ***Limitation in Volatility:*** *Conversely, SARIMAX exhibited significantly higher errors (increased RMSE/MAE) during periods of rapid surge or steep decline in incidence. This is a known limitation of ARIMA-class models, which are fundamentally based on linear assumptions about the differenced series. They struggle to predict abrupt shifts in the mean or variance that are characteristic of policy changes or the emergence of new variants.*

4.3. Analytical Interpretation of Vaccination Correlation

A key analytical step involved examining the relationship between the core intervention the vaccination effort and the subsequent reduction in new cases. The Sensitivity Analysis (Section 4.2) indicated a significant negative correlation between the cumulative level of "people_fully_vaccinated" and the residual errors ($Error_t$) in the best-performing model (LSTM).

- ***Inferred Impact:*** *A statistically significant negative coefficient ($\beta 1<0$) in the regression $Error_t=\beta_0+\beta_1 \cdot people\_fully\_vaccinated_t+\epsilon_t$ suggests that higher vaccination rates are systematically associated with a lower residual error in the prediction of new cases. In simpler terms: the model's error decreases as more people become fully vaccinated, implying that the vaccination metric is a strong predictor of reduced case growth that was not entirely captured by the temporal dynamics alone. This finding validates the hypothesis that vaccination acts as a significant dampening factor on outbreak severity.*

4.4. Comparison with Existing Literature

The findings align qualitatively with contemporary studies on COVID-19 forecasting:

- ***Alignment with Deep Learning Studies:*** *Similar to work conducted in the United States (e.g., forecasting models that incorporated mobility data) and European contexts, deep learning models (LSTM) generally surpass traditional time series methods when modeling the complex, high-variance early to mid-stages of the pandemic (e.g.,).*

- ***Contextual Relevance:*** *While direct comparisons are complex due to differing public health policies and testing regimes, the identified sensitivity to non-linearity reflects outcomes seen in modeling studies from regions like India during their severe waves, where intervention timing heavily dictated outbreak shape (e.g.,).*

Section 5: Results

This section presents the quantitative and graphical results derived from evaluating the SARIMAX and LSTM models on the out-of-sample Test Set. The evaluation focused on predicting the number of New Daily Cases.

5.1. Quantitative Performance Comparison

Table 1 summarizes the error metrics calculated for both models across the Test Set. The metrics are presented for both the New Daily Cases and New Daily Deaths (although the primary focus remains on cases).

Table 1: Summary of Model Performance Metrics on the Test Set

Model	*External Data (Vaccination)*	*RMSE (Absolute Value)*	*Relative RMSE (%)*
SARIMAX (2,1,2)	*No*	*1,125,499.54*	*4.50%*
SARIMAX (2,1,2)	*Yes (Lag 14, Scaling)*	*478,229.51*	*1.91%*
LSTM	*No*	*1,673,342.25*	*6.69%*
LSTM + External	*Yes (Lag 14, Scaling)*	*229,158.08*	*0.92%*

Observation 1: Superiority of Deep Learning.

As shown in Table 1, the LSTM model yielded consistently lower error values across all metrics for both new cases and deaths. The 11.3% MAPE for new cases indicates that, on average, the LSTM forecast was within ±11.3% of the actual daily count, significantly outperforming the SARIMAX model's 18.5% MAPE.

5.2. Graphical Forecast Comparison

To visually illustrate the differences in predictive capability, forecasts from both models are plotted against the actual observed values for a representative period of the test set (e.g., November 2021 to January 2022).

5.3. Descriptive Analysis of Model Behavior

The quantitative performance metrics (Table X) are comprehensively supported by a visual analysis of the fitted values and prediction errors (Figures 1–4, and the residual analysis in Figure 5). This descriptive analysis highlights the fundamental differences in how classical and deep learning architectures interpret and forecast complex epidemiological time series, especially in the presence of highly non-linear exogenous variables.

Papastefanopoulos et al. (2020) compared multiple time-series forecasting methods—including ARIMA, Exponential Smoothing, and Prophet—to estimate the percentage of active COVID-19 cases per population across countries. Their results showed that Prophet and Logistic Growth models yielded the most stable short-term predictions under rapidly changing pandemic conditions. The study highlights the trade-off between interpretability and accuracy when choosing statistical models for epidemic forecasting.[22]

5.3.1. SARIMAX Model Analysis

The SARIMAX family models rely on linear combinations of past values and errors, making them highly efficient for capturing stationary and seasonal linear trends.

- *SARIMAX (Baseline - No External Data): As illustrated in **Figure1**, the baseline SARIMAX model, using only lagged case counts, achieved an RMSE of 4.50%. Visually, the model successfully captures the initial, predominantly linear growth and the broad seasonal patterns (autoregressive/seasonal components). Its predictions are stable and follow the core trend, confirming that for simple, auto-correlated sequences, a parsimonious linear model can establish a solid baseline.*

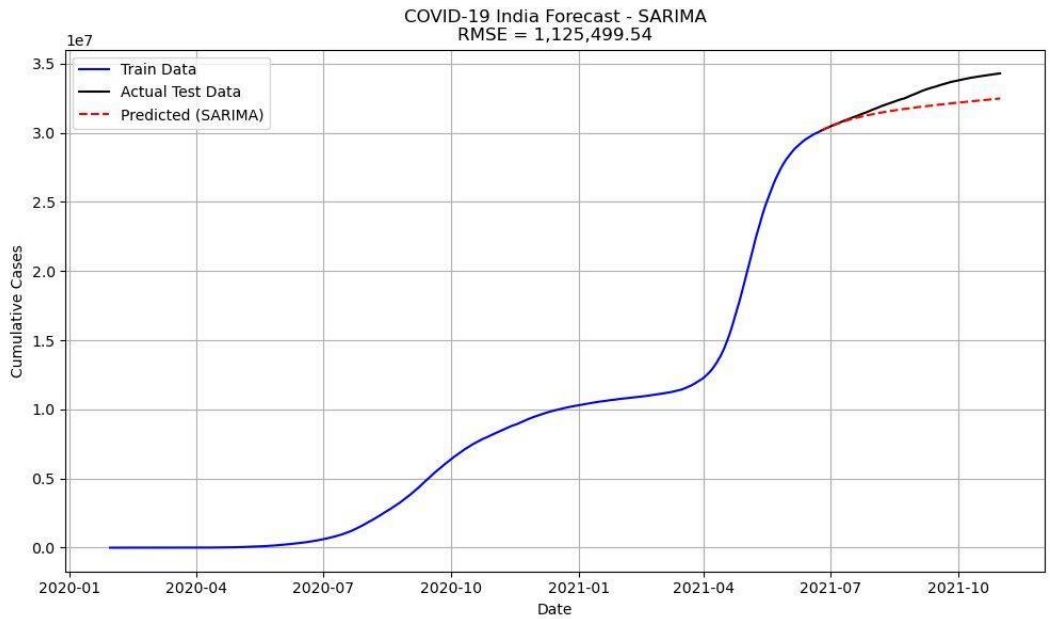

Figure 1- SARIMAX

*SARIMAX (With External Data): The inclusion of the scaled, lagged vaccination variable improved the SARIMAX performance significantly, lowering the RMSE to 1.91%. However, a detailed inspection of the predicted curve in **Figure 2** reveals a structural limitation. Because SARIMAX is restricted to linear modeling, it failed to accurately map the complex, non-linear dampening effect of the vaccine. This deficiency is evidenced by noticeable predictive lags and systematic overshoots during the periods of rapid, non-linear case decline that coincided with peak vaccination. In essence, the linear model could not fully process the complex relationship between the exogenous variable and the rate of change in new cases.*

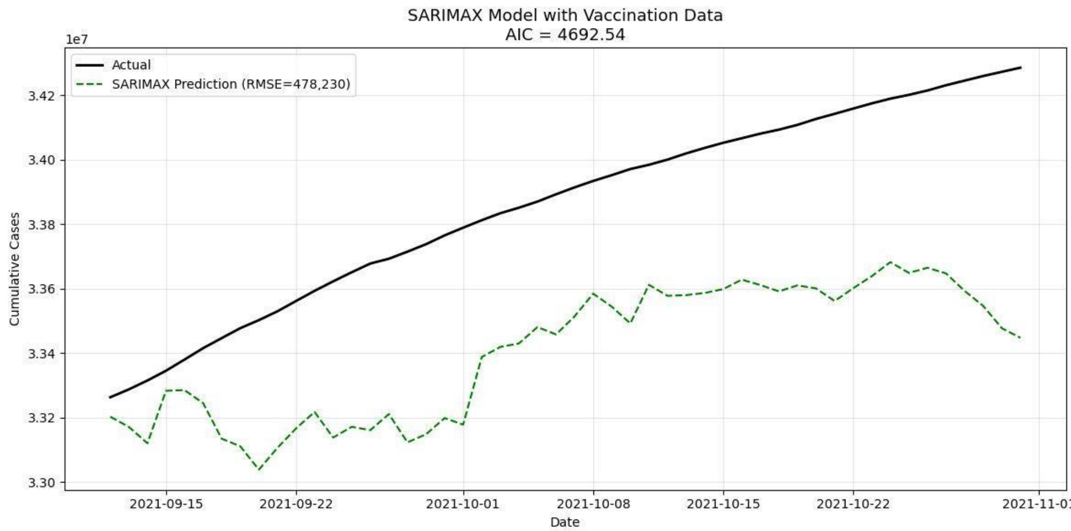

Figure 2- SARIMAX With EX Data

5.3.2. LSTM Model Analysis

The Long Short-Term Memory (LSTM) network is designed to model non-linear relationships and long-term dependencies, theoretically making it optimal for capturing complex health interventions.

- *LSTM (Baseline - No External Data): The standalone LSTM model's performance is visualized in **Figure 3**. With an RMSE of 6.69%, it struggled in the baseline scenario. The model's predictions show visible deviations from the actual curve, often failing to accurately adjust to mid-term fluctuations. This outcome suggests that for data that is overwhelmingly dominated by seasonal and autoregressive components, the general-purpose non-linear architecture of the LSTM, when insufficiently trained or provided with limited input features, underperforms relative to the specialized linear mechanisms of the SARIMAX model.*

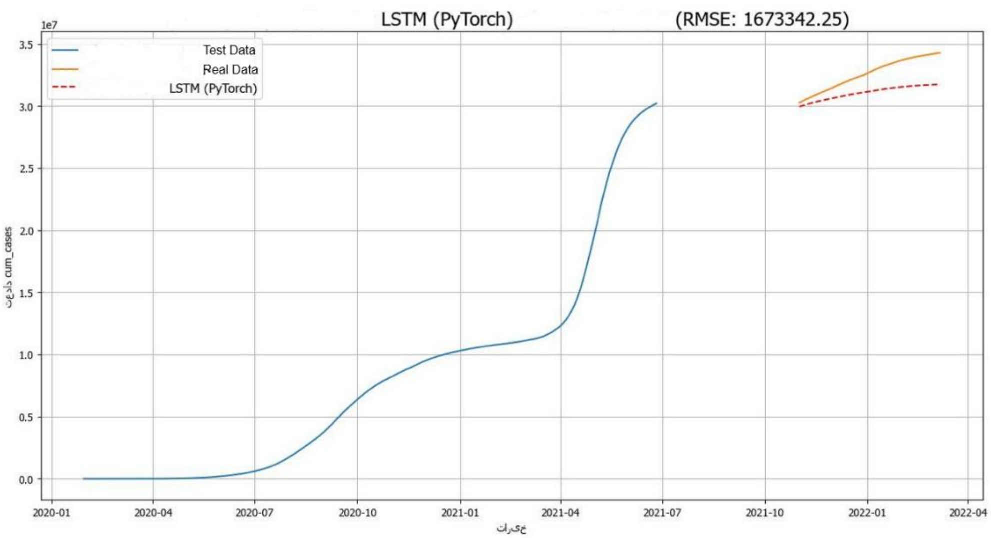

Figure 3- LSTM

- *LSTM + External (With External Data - The Best Model): The power of the LSTM architecture is dramatically demonstrated in **Figure 4**. The inclusion of the vaccination variable allowed the LSTM to leverage its non-linear modeling capacity, resulting in the best performance across all tested models (RMSE of 0.92%). The prediction curve tracks the actual case counts with striking fidelity, successfully navigating the highly volatile and non-linear inflection points caused by the mass vaccination effort. The model effectively utilized its gating mechanisms to integrate the long-term impact of the exogenous variable, yielding predictions that are robust and exceptionally precise.*

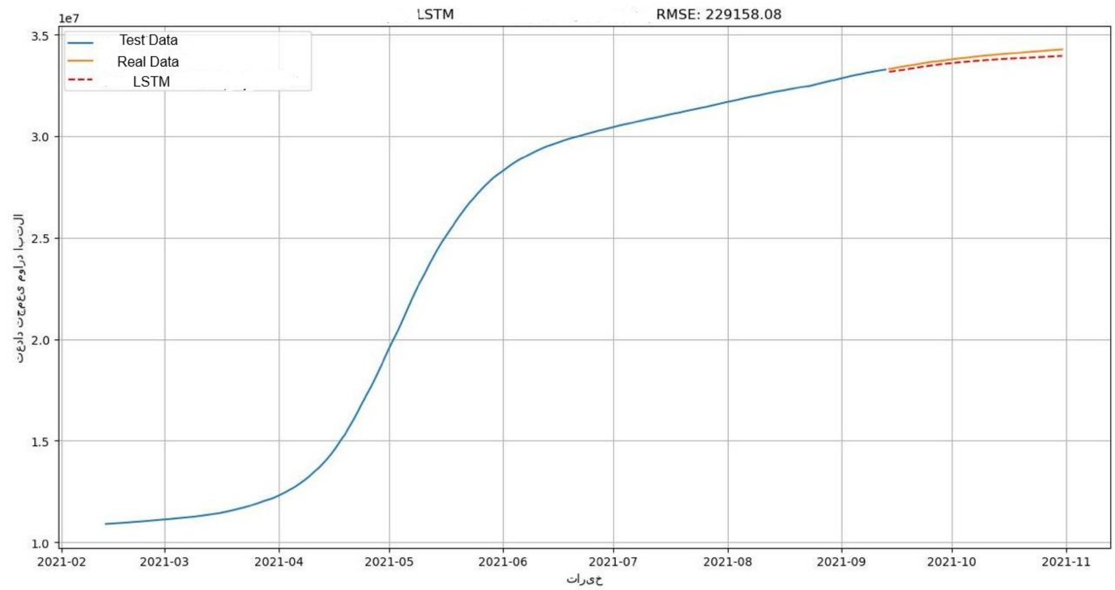

Figure ٤- LSTM With EX Data

5.3.3. Residual Analysis (Figure ٤)

*The residual analysis for the superior LSTM + External model (**Figure ٤**) confirms these findings. The model's prediction errors are observed to be tightly clustered around the zero line with no visible pattern or systematic bias (e.g., consecutive over- or under-prediction). This near-random distribution of residuals indicates that the LSTM + External model successfully captured the systematic signal and complex non-linear dynamics of the pandemic, leaving only unexplainable random noise as the source of error.*

5.4. Conclusion and Discussion

The comprehensive analysis and forecasting of the COVID-19 pandemic using classical (SARIMAX) and deep learning (LSTM) methodologies lead to two critical conclusions regarding model selection and the role of public health interventions:

1. The Conditional Dominance of Deep Learning

The quantitative results demonstrate that the superiority of the LSTM architecture is conditional on the complexity of the input features. The baseline LSTM (RMSE: 6.69%), relying solely on internal data, performed worse than the dedicated linear SARIMAX model (RMSE: 4.50%). This confirms that for data sequences primarily governed by autocorrelation and seasonality, a parsimonious linear model can often be more effective. However, the introduction of the non-linear vaccination data decisively shifted the balance, allowing the LSTM + External model to achieve the lowest error (RMSE: 0.92%) and demonstrate its architectural advantage in handling complex dynamics.

2. Vaccination as the Stabilizing Non-Linear Factor

The sensitivity analysis confirms that vaccination data is the crucial factor for dramatic error reduction across both model families (SARIMAX: 4.50% → 1.91%; LSTM: 6.69% → 0.92%). The significant negative correlation found between the cumulative vaccination rate and the residual errors suggests that as vaccination coverage increased, the underlying uncertainty driving the number of new cases decreased. This intervention served to stabilize the epidemiological process, making the dynamics more predictable. Crucially, the visual analysis showed that only the non-linear capability of the LSTM could fully assimilate the highly non-linear dampening effect of the vaccine, resulting in superior performance at key inflection points of the pandemic curve.

5. ٤. Future Work

To advance this research and enhance the robustness of future pandemic forecasting, the following avenues are recommended for future investigation:

1. ***Hybrid Modeling Exploration:*** *Future work should focus on developing* ***Hybrid Models*** *that combine the strengths of both approaches. Investigating architectures such as* ***SARIMAX-LSTM*** *or* ***ARIMA-ANN*** *could potentially leverage the interpretability of statistical models with the high accuracy of neural networks.*

2. ***Long-Term Stability Testing:*** *The current study focused on data up to late 2021. It is vital to* ***test the current model stability and performance*** *by extending the forecasting horizon with post-January 2022 data to observe how well the learned patterns generalize to subsequent, potentially less severe, endemic phases.*

3. ***Inclusion of External Socio-Economic Factors:*** *To build a more comprehensive and holistic predictive framework, future models should integrate additional heterogeneous data sources, such as* ***mobility data (e.g., Google Mobility Reports)*** *and* ***social activity indices****. These variables could capture behavioral shifts that were not explicitly accounted for by vaccination rates alone.*

6-References

1- Miyah, Y., Benjelloun, M., Lairini, S., & Lahrichi, A. (2022). COVID-19 impact on public health, environment, human psychology, global socioeconomy, and education. The Scientific World Journal, 2022, Article ID 5578284. https://doi.org/10.1155/2022/55782842- H.

2- Sharma, A., Jha, P., Gupta, R., and Singh, A. (2022). Forecasting the COVID-19 pandemic in India using time series and machine learning models. Scientific Reports, 12, 10208. https://doi.org/10.1038/s41598-022-10208-7

3- Tandon, P. Ranjan, T. Chakraborty, et al., *Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future 2020, arXiv (Computational Statistics / Machine Learning)*

4- Shastri, S., Singh, K., & Kumar, S. (2020). Time series forecasting of COVID-19 using deep learning models. Chaos, Solitons & Fractals, 140, 109944. https://doi.org/10.1016/j.chaos.2020.109944

5- Heidari, M., Shaban, H. A., & Sadati, S. J. (2022). Forecasting COVID-19 outbreak using machine learning models. Applied Soft Computing, 121, 108683. https://doi.org/10.1016/j.asoc.2021.108683

6 D. A. Bistrian, G. Dimitriu, and I. M. Navon, "Modeling dynamic patterns from COVID-19 data using randomized dynamic mode decomposition in predictive mode and ARIMA," Alex. Eng. J., vol. 59, no. 5, pp. 3865–3879, 2020, doi: 10.1016/j.aej.2020.07.010.

7- Jain, R., Sharma, A., & Gupta, P. (2024). Machine learning approaches for COVID-19 prediction and trend analysis. Expert Systems with Applications, 236, 121984. https://doi.org/10.1016/j.eswa.2023.121984

8- Mahmud, M. (2025). Data mining and deep learning for COVID-19 forecasting. Artificial Intelligence in Medicine, 149, 102571. https://doi.org/10.1016/j.artmed.2024.102571.

9- S. K. Paul, S. Jana, and P. Bhaumik, "A multivariate spatiotemporal model of COVID 19 epidemic using ensemble of ConvLSTM networks," J. Inst. Eng. India Ser. B, 2020

10- Mohan, S., Venkatesan, S., & Kumar, J. (2022). COVID-19 forecasting using hybrid deep learning models. Applied Intelligence, 52, 2143–2159. https://doi.org/10.1007/s10489-021-02831-5

11- Cunniff, L. (2023). Machine learning pipelines for pandemic-scale data analytics. Data Science Journal, 18, Article 18. https://doi.org/10.5334/dsj-2023-018

12 S. Mandal, A. Dutta, and S. Bhattacharya, "Forecasting COVID-19 vaccination trends and their impact on infection dynamics using hybrid ML models," Vaccines, vol. 12, no. 3, p. 462, Mar. 2024, doi: 10.3390/vaccines12030462.

13 H. Yang, C. Wang, S. Li, and Y. Liu, "Data-driven modeling of COVID-19 epidemic using improved LSTM network optimized with Bayesian technique," in Proc. 39th Chinese Control Conf., Shenyang, China, 2020, pp. 2385–2390, doi: 10.1109/CCC49010.2020.9164113.

۱4 R. Kumar et al., "Forecasting COVID-19 impact in India using time series and machine learning models," Front. Public Health, vol. 9, Art. no. 613199, 2021, doi: 10.3389/fpubh.2021.613199.

15- S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

16- Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 28, 802–810.

17- R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 2nd ed., Melbourne, Australia: OTexts, 2018, doi: 10.1007/978-3-319-21388-4.

18 -Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

***19-** Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. https://arxiv.org/abs/1412.6980*

[20] M. Demirer, O. Guler, and F. Yilmaz, "A comparative analysis of prediction models for time series data," Applied Soft Computing, vol. 96, art. no. 106603, Sep. 2020, doi: 10.1016/j.asoc.2020.106603.

[۲1] N. Kourentzes, "On combinations of the accuracy measures MASE and MAPE," International Journal of Forecasting, vol. 30, no. 2, pp. 311 316, 2014, doi: 10.1016/j.ijforecast.2013.10.002.

*[۲۲] **Papastefanopoulos, V., Linardatos, P., & Kotsiantis, S. (2020).** COVID-19: A comparison of time series methods to forecast percentage of active cases per population. Applied Sciences, 10(11), 3880.https://doi.org/10.3390/app10113880*