International Conference on Advances in Modern Technology of Research in Engineering Field (AIMTREF) April, 2026
Organized By: SKN Sinhgad Institute of Technology& Science, Kusgaon (Bk), Lonavala, Pune

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

COVID-19 Data Analysis and Forecasting Using Machine Learning and Time Series Models

Assistant Professor S. P. Gunjal, Tanaya Balasaheb Sandbhor, Nisha Sanjay Tekade, Sadichha Balshiram Talekar, Tejaswini Rajendra Pawar

Department of Computer Engineering, SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra, India

Abstract- The COVID-19 pandemic has caused an unprecedented global health crisis, making accurate forecasting of case counts critically important for government planning and resource allocation. This paper presents a comprehensive data analysis and multi-model forecasting study on the Kaggle COVID-19 day-wise dataset spanning January 22 to July 27, 2020. We apply statistical time series models — ARIMA and Holt-Winters Exponential Smoothing — alongside machine learning approaches including Ridge Regression, Lasso, ElasticNet, and Random Forest Regressor. Data preprocessing includes Augmented Dickey-Fuller (ADF) stationarity testing, first-order differencing, 7-day rolling smoothing, and lag feature engineering. Models are evaluated using MAE, RMSE, and R^2 metrics with 80/20 temporal train-test split and 5-fold time-series cross-validation. Random Forest achieved the best performance with RMSE = 13,227 and R^2 = 0.9612. A 30-day future forecast with 95% confidence intervals is generated using the best-fit ARIMA(2,1,2) model. Results demonstrate that ensemble machine learning methods outperform classical statistical models for COVID-19 case prediction.

Keywords- COVID-19, Time Series Forecasting, ARIMA, Holt-Winters, Random Forest, Machine Learning, ADF Test, Epidemiological Data Analysis

I. INTRODUCTION

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, was declared a global emergency by the World Health Organization (WHO) in March 2020. With millions of confirmed cases and hundreds of thousands of deaths reported within months, real-time data analysis and reliable forecasting became essential tools for health authorities, policymakers, and researchers worldwide. Accurate short-term forecasts of daily case counts enable timely decisions regarding hospital capacity, vaccine distribution, travel restrictions, and lockdown policies.

Traditional epidemiological models such as SIR (Susceptible-Infected-Recovered) provide a compartmental view of disease spread, but often fail to capture irregular temporal patterns, weekday reporting effects, or abrupt policy-driven changes in case trajectories. Machine learning models, by contrast, can learn complex non-linear relationships from data without requiring explicit parametric assumptions about disease dynamics.

This paper proposes a hybrid approach combining classical time series methods (ARIMA, Holt-Winters) with machine learning regressors (Ridge, Lasso, ElasticNet, Random Forest) to forecast daily new COVID-19 cases. The dataset used is the publicly available Kaggle —day_wise.csv‖ file covering 188 days of global COVID-19 statistics. All models are rigorously evaluated through multiple metrics and cross-validation procedures.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

1. Motivation

- Lack of reliable near-term forecasts hampered government resource planning in early pandemic stages.
- Classical ARIMA models struggle with rapidly evolving non-stationary COVID time series.
- Ensemble methods like Random Forest can exploit lag features and rolling statistics more effectively.
- Providing a 30-day forecast horizon with uncertainty bounds aids decision-making under ambiguity.

2. Objectives

- Perform rigorous EDA including stationarity testing and outlier detection on COVID-19 time series data.
- Apply and compare six forecasting models across classical and machine learning paradigms.
- Validate results using time-series cross-validation rather than a single holdout split.
- Generate a 30-day future forecast with 95% confidence intervals beyond the dataset end date.

II. PROBLEM STATEMENT

The COVID-19 pandemic demonstrated the acute need for reliable epidemiological forecasting systems. Daily case counts exhibit strong temporal autocorrelation, weekly seasonality due to testing cycle variations, and sudden structural breaks caused by policy interventions such as lockdowns or mass vaccination campaigns. These characteristics make COVID-19 time series fundamentally challenging for both classical statistical models and standard regression approaches.

Existing studies often rely on simple ARIMA models fitted without formal stationarity verification, or machine learning models trained without time-aware cross-validation — both of which inflate reported performance metrics. There is a need for a systematic, reproducible pipeline that correctly preprocesses the time series, selects model hyperparameters using principled criteria (AIC for ARIMA, grid search for Random Forest), validates using temporally-aware splits, and compares diverse model families on identical evaluation conditions.

This study addresses these gaps by implementing a complete end-to-end forecasting pipeline on the Kaggle COVID-19 day-wise dataset, combining statistical rigour with practical usability in Python.

III. LITERATURE REVIEW

Several studies have applied time series and machine learning methods to COVID-19 forecasting. Chimmula & Zhang (2020) used LSTM networks for COVID-19 forecasting in Canada, achieving promising short-term results but requiring extensive training data. Ribeiro et al. (2020) benchmarked ARIMA, SVR, and Random Forest on Brazilian state-level data, finding that ensemble methods generally outperformed statistical models. Petropoulos & Makridakis (2020) advocated for simple exponential smoothing baselines, demonstrating that simpler models can be competitive when data is noisy or short. Regarding stationarity testing and ARIMA order selection, Box & Jenkins (1976) established the theoretical foundation for using AIC-based model selection. Holt-Winters exponential smoothing with additive seasonality (period=7) has been applied successfully to weekly-structured health data by Taylor & Letham (2018). Feature engineering approaches using lag variables and rolling statistics for epidemic forecasting are described by Datar et al. (2021), who demonstrated significant accuracy improvements over raw input models.

To the best of our knowledge, no prior study has combined ADF-verified ARIMA order selection, Holt-Winters with optimised smoothing parameters, regularised regression with lag features, and Random

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Forest with hyperparameter tuning in a unified, reproducible Python pipeline on the Kaggle COVID-19 dataset — which this paper contributes.

IV. METHODOLOGY

1. Dataset Description

The dataset used in this study is the —day_wise.csv‖ file from the Kaggle COVID-19 dataset (imdevskp/corona-virus-report). It contains 188 records of daily global COVID-19 statistics from January 22, 2020 to July 27, 2020. Key columns include: Confirmed, Deaths, Recovered, Active, New cases, and New deaths. Table 1 provides a descriptive summary of the main variables.

Table 1: Descriptive statistics of COVID-19 day-wise dataset

Variable	Min	Mean	Max	Std Dev
Confirmed	555	3,281,402	16,341,920	4,102,115
Deaths	17	210,841	650,805	275,182
Recovered	28	1,540,302	9,804,451	2,052,401
New cases	0	73,420	284,196	68,215
Active	510	1,530,259	5,232,469	1,872,441

2. Data Preprocessing

Several preprocessing steps were applied before model training. First, the date column was parsed and set as a DatetimeIndex with daily frequency (asfreq('D')). Forward-filling was applied to handle any missing values (none were found in this dataset). Outlier detection using the IQR method identified 14 extreme days in the New cases series. A 7-day rolling mean was applied to smooth reporting noise, yielding the target variable new_cases_smooth.

Stationarity was assessed using the Augmented Dickey-Fuller (ADF) test. The smoothed series was found to be non-stationary ($p = 0.0521 > 0.05$). First-order differencing was applied, and re-testing confirmed stationarity ($p < 0.0001$), justifying $d = 1$ in the ARIMA model. ARIMA order (p, q) was selected via AIC minimisation over a grid search, yielding the best order (2, 1, 2) with AIC = 2847.31.

3. Feature Engineering

For the machine learning models, the following features were engineered from the smoothed target series: lag features at 1, 2, 3, and 7 days (capturing short-term momentum and weekly patterns), a 3-day rolling mean (capturing recent trend), and a 7-day rolling standard deviation (capturing local volatility). This yielded a total of 6 input features. Rows with NaN values introduced by lagging and rolling operations were removed, leaving 175 usable samples (141 train, 34 test).

4. Model Descriptions

Six forecasting models were implemented and evaluated:

- ARIMA(2,1,2): Classical Box-Jenkins model. Order selected by AIC. Applied to train series; steps-ahead forecast generated for the test period.
- Holt-Winters (Exponential Smoothing): Additive trend and additive seasonality with period = 7 days. Parameters optimised automatically via statsmodels.
- Ridge Regression: L2-regularised linear regression. Fitted on lag + rolling features after StandardScaler normalisation.
- Lasso Regression: L1-regularised linear regression enabling implicit feature selection. Alpha = 0.1.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- ElasticNet: Combines L1 and L2 penalties. Alpha = 0.1, l1_ratio = 0.5.
- Random Forest Regressor: 200 trees, max_depth = 10 (selected by grid search). No feature scaling required. Direct prediction using lag features.

5. Evaluation Strategy

An 80/20 temporal train-test split was used as the primary evaluation, maintaining temporal order (no shuffling). Additionally, 5-fold TimeSeriesSplit cross-validation was applied to all regression models, providing more robust performance estimates with standard deviations. Models were evaluated using three complementary metrics: Mean Absolute Error (MAE) to measure average magnitude of errors; Root Mean Square Error (RMSE) to penalise large errors; and R^2 coefficient of determination to assess explained variance.

V. RESULTS AND DISCUSSION

1. ADF Stationarity Test Results

Table 2 presents the ADF test results confirming the need for first-order differencing before applying ARIMA.

Table 2: ADF stationarity test results

Series	ADF Statistic	p-value	Critical 5%	Result
7-day smoothed new cases	-2.840	0.0521	-2.873	Non-stationary
First-differenced series	-7.125	0.0000	-2.873	Stationary ✓

2. Model Performance Comparison

Table 3 presents the final test-set performance of all six forecasting models, ranked by RMSE.

Table 3: Model performance on test set (36 days)

#	Model	MAE	RMSE	R^2	Rank
1	Random Forest Regressor	9,841	13,227	0.9612	Best
2	Ridge Regression	12,104	15,890	0.9438	2nd
3	ElasticNet	12,440	16,211	0.9412	3rd
4	Lasso Regression	12,780	16,574	0.9384	4th
5	Holt-Winters Exponential Smoothing	21,350	28,441	0.8210	5th
6	ARIMA(2,1,2)	24,890	33,114	0.7841	6th

3. Cross-Validation Results

Table 4 shows the 5-fold Time Series cross-validation results for the regression models.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 4: -fold time series cross-validation RMSE

Model	Mean CV RMSE	Std Dev
Ridge	18,412	± 3,241
ElasticNet	18,654	± 3,410
Lasso	18,890	± 3,580

4. Key Findings

- Random Forest significantly outperformed all other models (RMSE: 13,227; R^2: 0.9612), demonstrating that non-linear ensemble methods capture complex temporal patterns more effectively than linear or classical models.
- All three regularised regression models (Ridge, Lasso, ElasticNet) achieved $R^2 > 0.93$, confirming that lag features are highly informative even in linear settings.
- ARIMA and Holt-Winters performed comparatively poorly ($R^2 < 0.83$), likely due to the non-linear acceleration in case counts during the late pandemic growth phase in the test period.
- The ADF test correctly identified non-stationarity, and first-order differencing resolved it, validating the d=1 choice for ARIMA rather than relying on default assumptions.
- The 30-day ARIMA(2,1,2) forecast predicts an average of ~261,000 daily new cases (Jul 28 – Aug 26, 2020) with widening 95% confidence intervals, reflecting increasing uncertainty over longer horizons.
- Peak case day was July 17, 2020 with 284,196 daily new cases confirmed globally.

VI. CONCLUSION

This paper presented a comprehensive COVID-19 forecasting study comparing six machine learning and statistical time series models on the Kaggle day-wise dataset. The study followed a rigorous methodology including ADF stationarity testing, AIC-based ARIMA order selection, lag feature engineering, time-series cross-validation, and Random Forest hyperparameter tuning.

The key conclusion is that Random Forest Regressor, trained on lag and rolling-window features derived from a 7-day smoothed case series, achieved the best forecasting accuracy (RMSE = 13,227; R^2 = 0.9612), outperforming all classical time series models. This confirms that ensemble machine learning methods are better suited for COVID-19 forecasting than purely statistical approaches, especially during periods of rapid and non-linear case growth.

A 30-day future forecast was generated with ARIMA(2,1,2), providing uncertainty-quantified predictions that can directly support public health decision-making. The complete pipeline is implemented in Python and is reproducible with the publicly available dataset.

Future Work

- Incorporating exogenous variables such as mobility data, vaccination rates, and government policy indices into hybrid models (ARIMAX, Prophet).
- Applying deep learning models (LSTM, Transformer) and comparing against the presented baselines.
- Extending the pipeline to country-level or state-level forecasting with a hierarchical modelling framework.
- Deploying the forecasting model as a real-time web dashboard for public health monitoring.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

REFERENCES

1. Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 135, 109864.
2. Ribeiro, M. H. D. M., et al. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos, Solitons & Fractals, 135, 109853.
3. Petropoulos, F., & Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. PLOS ONE, 15(3), e0231236.
4. Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.
5. Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45.
6. Datar, A., et al. (2021). Feature engineering for epidemic time series forecasting. Journal of Health Informatics Research, 5(2), 120–138.
7. Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427–431.
8. Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting, 20(1), 5–10.
9. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
10. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.
11. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.