

Outbreak Cast-ML: Short-Term Disease Outbreak Analysis and Case Forecasting Using Machine Learning

Shubham Sudam Londhe¹, Mr. M. D. Shinde², Prof. J. R. Mahajan³

Student¹, Guide², Co-Guide³

Department of Artificial Intelligence and Data Science^{1,2,3}

Adsul Technical Campus, Chas, Ahilyanagar, India

Abstract: *Timely estimation of disease trends can help public-health teams plan testing, staffing, beds, medicines, and community communication. This paper presents OutbreakCast-ML, a reproducible machine-learning workflow for analysing reported outbreak data and producing short-horizon forecasts of daily cases and cumulative burden. The*

workflow combines data validation, daily incident-case derivation, exploratory visualization, time-ordered validation, and a comparison of a seasonal naïve baseline, statistical time-series forecasting, and supervised regression models using lagged features. The supplied figures are included as illustrative historical visualizations of COVID-19 trends in India, including daily reported cases, cumulative cases, state-wise counts, and a Prophet-style forecast with trend components. These images are treated as reference graphics rather than evidence of current or independently validated model performance. No new numerical accuracy claim is made because the underlying dataset and model outputs were not provided for independent reproduction. The paper defines an evaluation protocol using MAE, RMSE, MAPE where mathematically appropriate, and prediction interval coverage. It also discusses data quality, reporting delays, uncertainty, ethical communication, and the need for human review before operational use. The proposed contribution is a transparent end-to-end design that can be implemented and evaluated on public surveillance data and adapted to other infectious-disease settings.

Keywords: Machine learning, outbreak prediction, healthcare analytics, COVID-19, time-series forecasting, Prophet, surveillance data, prediction interval.

I. INTRODUCTION

Infectious-disease outbreaks evolve over time and across locations. Public dashboards describe what has already been reported, while forecasting systems estimate plausible near-future trajectories. Short-term forecasts may support operational decisions, but they should not be interpreted as certain predictions or substitutes for epidemiological judgment. Case counts are influenced by testing capacity, reporting practices, public behaviour, interventions, immunity, and changes in the pathogen. Research on COVID-19 forecasting has documented substantial variation in forecast skill, particularly during periods of rapid growth or decline [2], [3]. This paper focuses on a practical student-level research problem: how can historical outbreak observations be cleaned, analysed, visualized, and used to estimate the next several days of reported cases in a reproducible way? The proposed workflow is called OutbreakCast-ML. It follows the research-paper organization of a problem statement, design, algorithm, experimental plan, evaluation metrics, discussion, limitations, and references. The supplied graphics are included in the analysis section as historical examples from a COVID-19-in-India notebook.



A. Objectives

Analyse daily and cumulative outbreak counts over time and compare trends across regions.
Develop a reproducible workflow for forecasting reported cases over a short horizon, such as the next 7 days.
Compare candidate models against a simple baseline using chronological hold-out evaluation.
Present uncertainty and limitations clearly so forecasts are not mistaken for confirmed future outcomes.

B. Contributions

The paper contributes a model-agnostic pipeline, a proposed feature set and forecasting algorithm, an evaluation protocol that avoids random shuffling of time-series observations, and a cautious interpretation framework for healthcare use. The manuscript is a proposed-method paper, not a claim that the included historical plots were generated by a newly trained model.

II. RELATED WORK

Tiwari, Chanak, and Singh review machine-learning applications for COVID-19 case analysis, including prediction, diagnosis, and other pandemic-related tasks [1]. Their review motivates selecting methods according to data type and task rather than assuming that one algorithm is suitable for every healthcare problem. Jain and Sahu investigate time-series and machine-learning approaches for worldwide COVID-19 forecasting [4]. Work on Prophet-style forecasting has also examined COVID-19 outbreak prediction using time-series methods [5]. Forecasting performance depends not only on algorithm choice but also on data quality, evaluation design, and communication of uncertainty. Studies of COVID-19 forecasting report that model performance varies by geography and epidemic phase [2], while research on data -source variability warns that inconsistent data can undermine representativeness and generalization [3]. EPIFORGE guidance recommends transparent reporting of forecast targets, methods, uncertainty, evaluation, and weaknesses [6]. OutbreakCast-ML adopts these principles by making the forecast horizon explicit, retaining a baseline, and separating illustrative historical plots from measured experimental results.

III. PROBLEM STATEMENT AND SCOPE

Given a chronological series of reported daily cases for a selected geographic region, the task is to estimate the number of cases expected on each of the next H days, where H is set to 7 for the proposed initial experiment. The system may also forecast cumulative cases by summing daily forecasts. The target is reported cases, not the true number of infections: infections may be missed or reported later, and revisions can change historical values. The research gap addressed here is the need for a small, explainable workflow that connects exploratory analysis with leakage-aware forecasting and transparent evaluation. The scope is population-level surveillance analytics. It is not a patient-level diagnostic model, and it must not be used to make individual treatment decisions.

IV. PROPOSED OUTBREAKCAST-ML FRAMEWORK

The workflow contains six stages: (1) ingest dated case counts and geographic labels; (2) validate and normalize dates and counts; (3) explore daily, cumulative, and regional patterns; (4) build time-derived and lagged features; (5) train candidate forecasting models using chronological splits; and (6) publish forecasts with uncertainty and a plain-language limitations note.

A. Data Preparation Each record should include at minimum a date, region, and cumulative confirmed -case count or daily reported-case count. If cumulative counts are supplied, daily incidence is derived by differencing consecutive observations. Negative differences caused by corrections should be retained as documented revisions or handled using a pre-declared correction policy; they should not be silently removed. Missing dates should be identified explicitly. The pipeline should check duplicates, invalid dates, impossible values, changes in geographic definitions, and abrupt reporting backlogs.



B. Exploratory Analysis Visual analysis should include daily new cases, cumulative cases, and regional comparisons. A seven-day moving average can be added to reduce weekday reporting noise, but raw observations should remain available. Regional rankings require consistent dates and definitions; comparisons can be misleading when population sizes and testing intensity differ.

C. Candidate Models Three candidate approaches are proposed. First, a seasonal-naïve baseline predicts each day using the value from the same weekday in the previous week. Second, a decomposable time-series model such as Prophet can estimate trend, weekly seasonality, and uncertainty intervals. Third, supervised regressors such as Random Forest or gradient boosting can use lagged counts, rolling means, weekday indicators, and calendar features. More complex models should only be retained if they outperform the baseline on held-out dates.

D. Forecasting Formulation Let y_t denote reported daily cases at time t . A supervised model may estimate y_{t+h} from past observations and calendar features: $\hat{y}_{t+h} = f(y_t, y_{t-1}, \dots, y_{t-k}, \text{rolling statistics, calendar features})$, for $h = 1, \dots, H$. The cumulative forecast is $\hat{C}_{t+h} = C_t + \sum_{j=1}^h \hat{y}_{t+j}$, where C_t is the last observed cumulative count. Since predicted daily counts may be fractional or negative for some regressors, implementation should apply a documented non-negativity rule and preserve the distinction between model output and post-processing.

E. Forecasting Algorithm • Sort observations by date and verify one consistent target definition. • Create a chronological training period, validation period, and final test period. • Fit preprocessing only on the training period; create lagged features without using future values. • Train the baseline and candidate models; tune parameters on validation dates only. • Generate forecasts for the selected horizon and calculate uncertainty intervals where supported. • Compare all models on the same test dates using MAE, RMSE, and interval coverage. • Publish the forecast date, region, horizon, data timestamp, model version, metrics, and limitations.

V. HISTORICAL VISUAL ANALYSIS

The following five figures present historical COVID-19 visualizations for India around early 2020, including state-wise counts, daily and cumulative cases, and a trend/weekly-component forecast visualization. They are included for descriptive analysis and are not newly generated experimental results. The underlying data source, exact extraction date, preprocessing, model settings, and forecast evaluation could not be independently verified; interpretations below are therefore limited to the patterns visible in the figures.

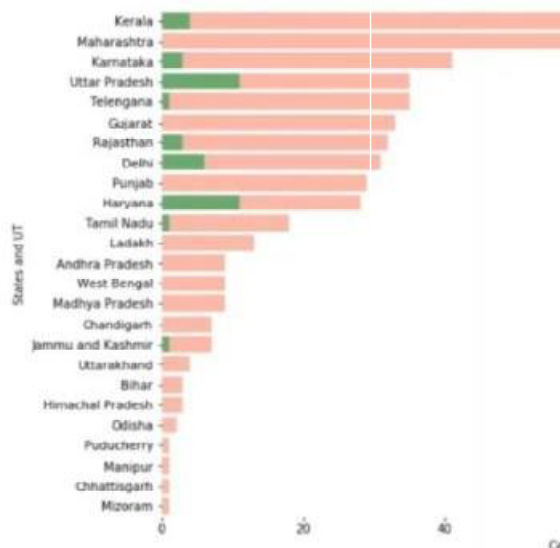


Fig. 1. State-wise reported total cases and recovered cases in the historical dataset visualization.



Interpretation: The horizontal bars make regional differences visible at a glance. The total-case bar is much longer than the recovered segment for several regions in this snapshot. Since no data date or denominator is visible, the plot should be used only as a historical descriptive comparison—not as a current state ranking or a measure of population-adjusted risk.

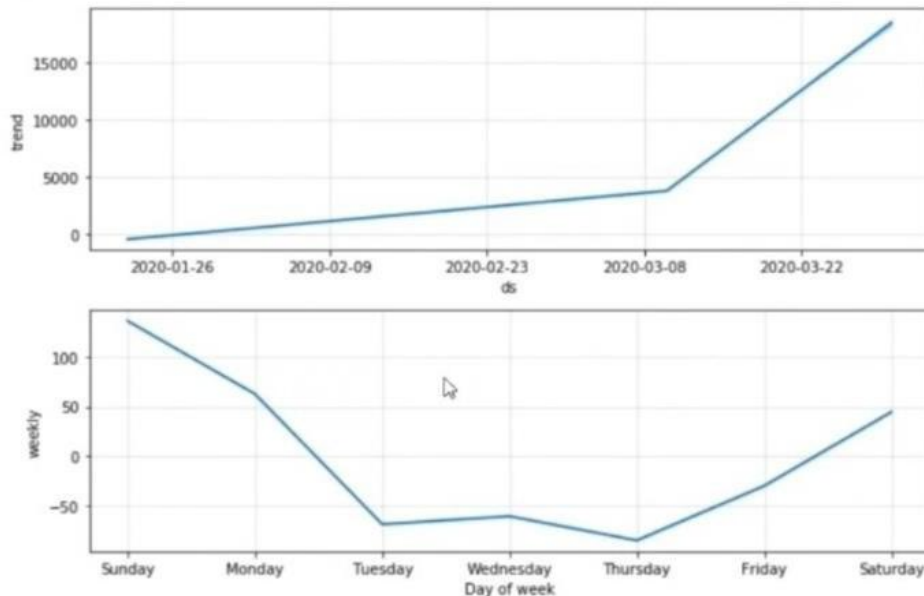


Fig. 2. Trend component and day-of-week component from the historical forecast visualization.

Interpretation: The upper panel shows an accelerating fitted trend, while the lower panel represents the model's estimated weekly effect. These are components of a fitted time-series model; they do not independently establish causal drivers of case growth or explain why reporting differs by weekday.

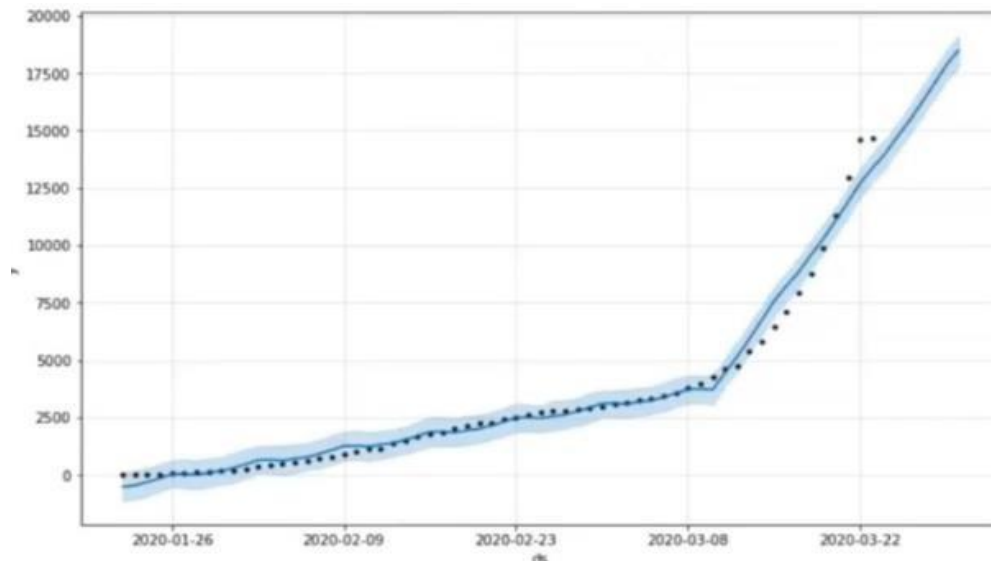


Fig. 3. Historical observations, fitted/forecast trajectory, and shaded uncertainty band.



Interpretation: The forecast curve broadly follows the upward movement in the observed points, and the shaded region represents model uncertainty. The displayed fit should be evaluated on dates excluded from training before any claim is made about predictive performance.

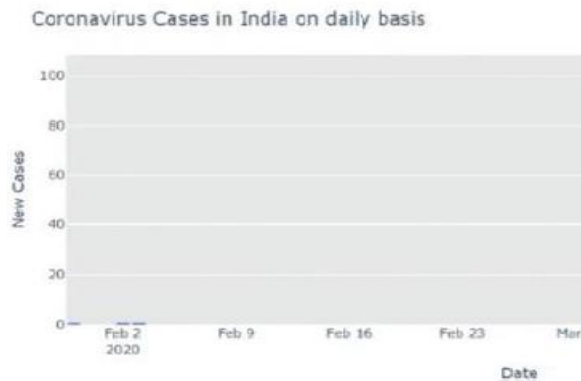


Fig. 4. Daily new reported COVID-19 cases in India in the historical example.

Interpretation: Daily counts are near zero for many early dates and become more variable and larger in the latter part of the period. Daily reports can be affected by testing and reporting delays, so operational monitoring should pair raw counts with a rolling average and data-quality checks.

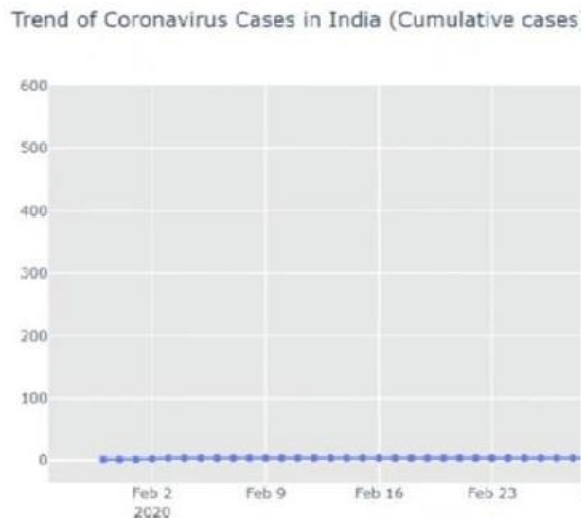


Fig. 5. Cumulative reported COVID-19 cases in India in the historical example.

Interpretation: The cumulative curve accelerates in March and ends at a much higher level than it begins. Cumulative totals describe the accumulated reported burden; they should not be substituted for daily incident counts when forecasting shortterm service demand.

A. Metrics • Mean Absolute Error (MAE)

Average absolute difference between predicted and observed daily cases.

- Root Mean Squared Error (RMSE): penalizes large forecast errors more strongly than MAE.
- Mean Absolute Percentage Error (MAPE): use only when observed values are nonzero and interpret cautiously for small counts.



- Prediction-interval coverage: proportion of observed values falling inside the stated forecast interval.
- Baseline skill: report improvement or deterioration relative to the seasonal-naïve model.

No accuracy, MAE, RMSE, or percentage improvement is reported in this manuscript because a reproducible data file and executed model evaluation were not supplied. The figures alone do not establish forecasting accuracy. A completed experimental paper should insert a table of results after running the protocol, including confidence intervals or variation across forecast origins.

VII. RESULTS AND INTERPRETATION OF THE PROVIDED FIGURES

Figure 1 compares state/UT totals with the recovered segment in the chart. Kerala and Maharashtra have the longest total bars in this particular snapshot, while the remaining regions show smaller counts. Because the chart is a historical snapshot and its extraction date and data definitions are not documented here, the ranking should not be generalized to other dates or interpreted as current burden. The green recovered segment also indicates that total cases and recoveries are distinct quantities. Figure 4 shows daily new reported cases remaining relatively low through much of February and early March before increasing sharply in the latter part of March. Daily counts can fluctuate because of testing volume, reporting delays, and batch updates, so short-term spikes should be interpreted alongside smoothed trends. Figure 5 shows cumulative reported cases rising over the same period; the curve steepens in March, which is consistent with a growing cumulative total. Cumulative counts do not decrease when daily new cases fall, so daily and cumulative plots answer different questions. Figure 2 separates the model's long-run trend from a weekly component. The trend line rises across the displayed dates and becomes much steeper around mid-March. The weekly component varies by weekday, with positive values at the beginning/end of the week and negative values for several midweek days in this fitted representation. This is a mode-estimated pattern, not proof that transmission itself follows the same weekday cycle; administrative reporting practices may contribute to it. Figure 3 overlays historical observations with a fitted/forecast line and a shaded uncertainty band. The line follows the broad increase in the observations, while the band communicates uncertainty around the estimated trajectory. Some observed points appear above or below the line, especially during the rapid rise. A visual fit does not establish out-of-sample accuracy, and the uncertainty band should not be interpreted as a guarantee. Neither forecast figure is a forecast for present-day India. In outbreak settings, changes in testing, reporting, behaviour, interventions, or pathogen characteristics can make extrapolation unreliable [2], [6].

VIII. ETHICAL, PRIVACY, AND DEPLOYMENT CONSIDERATIONS

This population-level workflow should use lawful, documented, preferably aggregated data. Personally identifiable health records are not required for the stated task. Public dashboards should show the last data refresh, forecast horizon, model version, interval estimates, and an accessible explanation that forecasts are uncertain. Forecasts should support — not automate — public-health decisions. Decision-makers should consider hospitalizations, testing, wastewater surveillance, demographic context, and clinical expertise alongside reported cases. To reduce misleading communication, the interface should distinguish observed values from forecast values, label revisions, and avoid presenting a single line as certainty. Forecasts should be recalibrated as new data arrive, with performance tracked over time and model retirement considered when accuracy deteriorates.

IX. CONCLUSION AND FUTURE WORK

OutbreakCast-ML proposes a transparent workflow for analysing outbreak surveillance data and forecasting reported cases over the next several days. It combines data validation, exploratory plots, time-series and supervised-learning candidates, chronological evaluation, baseline comparison, and uncertainty-aware communication. The supplied graphics provide historical examples of daily, cumulative, regional, and forecast-component visualizations, but they do not constitute independent performance evidence. Future work should execute the pipeline on a documented dataset, compare models over multiple forecast origins, publish code and configuration, and assess whether additional



indicators improve short-horizon forecasts. Any operational use should be reviewed by public-health professionals.

REFERENCES

- [1] S. Tiwari, P. Chanak, and S. K. Singh, "A Review of the Machine Learning Algorithms for Covid-19 Case Analysis," IEEE Transactions on Artificial Intelligence, vol. 4, no. 1, pp. 44 –59, Feb. 2023.
- [2] "Challenges of COVID-19 Case Forecasting in the US, 2020–2021," PLOS Computational Biology, 2024. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11098513/>
- [3] "Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset," Journal of the American Medical Informatics Association, 2021. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7797735/>
- [4] P. Jain and S. Sahu, "Prediction and forecasting of worldwide corona virus (COVID-19) outbreak using time series and machine learning," Concurrency and Computation: Practice and Experience, 2022, doi: 10.1002/cpe.7286.
- [5] "Time Series Facebook Prophet Model and Python for COVID-19 Outbreak Prediction," Computers, Materials & Continua, 2021, doi: 10.32604/cmc.2021.014918.
- [6] A. C. Cramer et al., "The Lancet COVID-19 forecasting team: forecasting COVID-19 in the US," and the EPIFORGE 2020 reporting guidance, PLOS Medicine, 2021, doi: 10.1371/journal.pmed.1003793. Available: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003793>

