Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 8 (2026) 3528

Journal of the Nigerian Society of Physical Sciences

An empirical evaluation of mathematical, machine learning, and hybrid modeling approaches for COVID-19 forecasting

Pooja Satwani[a], Sudhir Kumar Mishra[b,*]

[a]Department of Mathematics, Nims Institute of Engineering & Technology, Nims University Rajasthan, Jaipur, 303121, Rajasthan, India
[b] Department of Electronics & Communication Engineering, Nims Institute of Engineering & Technology, Nims University Rajasthan, Jaipur, 303121, Rajasthan, India

Abstract

Accurate forecasting of infectious disease spread is essential for public health decision-making. This study compares mechanistic susceptible-exposed-infectious-removed (SEIR) models, machine-learning methods, and hybrid artificial intelligence (AI)-driven epidemiological frameworks using coronavirus disease 2019 (COVID-19) case data from India, the United States, Italy, and Japan. After preprocessing and exploratory analysis, we estimate time-varying transmission rates and develop two hybrid extensions: one predicts $\beta(t)$ using machine learning, and the other learns residual errors from the SEIR model. A baseline machine-learning model using lag-based and effective reproduction number (R_t)-driven features is also evaluated. Performance is assessed using mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), high-incidence MAPE (on days with observed incidence ≥ 100 cases), and peak-timing error. With the SEIR transmission-rate spline fitted strictly to training data and all models evaluated on an identical 223-day held-out test window, a simple persistence baseline outperforms the mechanistic SEIR model, both hybrid AI extensions (Tracks A and B), and the pure machine-learning Random Forest model across all four countries (Diebold–Mariano test, $p < 0.001$ in every comparison). The hybrid AI components do not consistently improve forecasting accuracy relative to the SEIR baseline. None of the mechanistic or machine-learning approaches anticipates new epidemic waves beyond the training period, underscoring the difficulty of medium-range epidemic forecasting under genuinely held-out conditions. The comparative evaluation highlights the complementary strengths and limitations of these approaches and provides guidance for future epidemic forecasting models.

DOI: 10.46481/jnsps.2026.3528

Keywords: SEIR model, Hybrid modeling, Infectious disease prediction, Epidemic forecasting, COVID-19

Article History:
Received: 17 May 2026
Received in revised form: 20 July 2026
Accepted for publication: 20 August 2026
Available online: 07 September 2026

Communicated by: B. J. Falaye

1. Introduction

Infectious diseases such as influenza, Ebola, severe acute respiratory syndrome (SARS), and coronavirus disease 2019 (COVID-19) continue to demonstrate the vulnerability of global health systems. These outbreaks underscore the critical importance of reliable forecasting methods. Mathematical modeling has proven instrumental in helping researchers understand how epidemics spread. One of the earliest approaches is the classical susceptible-infectious-removed (SIR) model proposed by [1]. Later, researchers developed extended models such as susceptible-exposed-infectious-removed (SEIR), which include a latent (exposed) stage of infection. Because of this feature, SEIR models are widely used in COVID-19 studies to estimate disease transmission rates and reporting delays across different regions [2–4] (note: Gupta *et al.* [2] is a preprint).

In many cases, traditional compartmental models assume

*Corresponding Author Tel. No.: +91-708-376-3152.
Email address: sudhir.mishra@nimsuniversity.org (Sudhir Kumar Mishra)

homogeneous population mixing and slowly changing model parameters, which can limit their ability to adapt during rapidly changing outbreaks. Empirical studies have demonstrated that real epidemics are influenced by behavioral changes, public-health interventions, mobility patterns, meteorological factors, and under-reporting—drivers that are difficult to capture using fixed-parameter formulations [5–7]. As a result, SEIR-type models often provide useful structural interpretation of epidemic processes but can struggle to maintain short-term predictive accuracy during multi-wave outbreaks [2, 5].

Complementary to mechanistic methods, machine-learning approaches have been successfully applied to epidemic forecasting and related tasks. Tree-based models such as XGBoost and Random Forests, as well as temporal architectures like LSTM, have been employed to capture nonlinear and temporal patterns in COVID-19 case series and to integrate auxiliary information for short-term prediction [6–8]. The systematic reviews indicate that machine learning techniques can achieve high predictive accuracy however their performance depends strongly on data quality and may not generalize across different geographical locations or epidemic phases. [5, 9].

Hybrid modeling frameworks combine mechanistic structure with data-driven learning to improve both interpretability and predictive flexibility. Some studies have used SEIR or SEIRD models corrected with ARIMA residuals [10], while others have integrated machine-learning methods within compartmental models [11]. Still others employ spatio-temporal pipelines that leverage real-world data to improve forecasting performance [12]. Hybrid approaches can capture mechanistic constraints while adapting to changing patterns in surveillance data [11, 12]. Similar hybrid mechanistic–machine-learning approaches have also been applied outside epidemiology, for example, in the prediction of cardiovascular disease [13], illustrating the broader applicability of this modeling paradigm rather than its direct relevance to the forecasting of COVID-19.

The existing studies continue to face several challenges as surveillance data often exhibit variations in reporting quality. Parameter non-identifiability can also make inference difficult when using confirmed case data. Many studies also focus on limited geographic regions or only short-term forecasting horizons [5, 9]. These limitations motivate the need for comparative studies that evaluate mechanistic models, machine-learning approaches and hybrid frameworks across different settings using reliable evaluation metrics. This study addresses these gaps through a multi-country comparative evaluation of forecasting methods using COVID-19 case data from India, the United States, Italy, and Japan. The first step estimates time-varying transmission rates within a classical SEIR framework using smoothed incidence data. After that two hybrid extensions are developed which integrate machine-learning components into the mechanistic model. A purely data-driven forecasting model is also used as a benchmark.

Instead of treating artificial intelligence as a replacement for mechanistic epidemic models this work considers AI as a complementary predictive layer that can improve adaptability under noisy and non-stationary surveillance conditions. The aim is not only to maximize predictive accuracy. The goal is also to understand how AI methods can be systematically integrated within epidemiological modeling frameworks to improve short-term responsiveness, uncertainty estimation, and robustness across different epidemic settings.

By comparing structural, hybrid, and machine-learning-based forecasts across several countries and epidemic conditions, this study aims to clarify the strengths and limitations of each modeling approach. The findings also provide practical guidance for building adaptive and interpretable epidemic forecasting systems.

While several prior studies have proposed hybrid mechanistic–machine-learning frameworks for epidemic forecasting (e.g., SIMLR [11], SEIRD–ARIMA corrections [10], spatio-temporal attention networks [12]), most report favourable outcomes for the hybrid approach on a limited number of countries or epidemic phases. The present study differs in three respects. First, it evaluates two distinct hybrid pathways (a machine-learned transmission rate and a machine-learned residual correction) alongside a purely data-driven benchmark and a simple persistence baseline, across the same four countries, under an identical evaluation framework, rather than reporting a single hybrid variant in isolation. Second, it reports a finding that is comparatively under-represented in the literature: under the surveillance conditions and countries studied here, the hybrid extensions did not consistently improve forecasting accuracy relative to the mechanistic baseline, though the effect of residual learning (Track B) was smaller and more consistent than initially expected. Third, it adopts a strictly training-only parameter-fitting procedure for the SEIR baseline and its derived hybrid components, avoiding the partial information leakage into Track A's reproduction-number feature that can arise from full-period fitting and that is often left unaddressed in comparative hybrid-modeling studies. Taken together, these contributions provide a more methodologically transparent and cautionary perspective on when and why hybrid AI-epidemiological models succeed or fail, complementing the predominantly positive results reported elsewhere in the literature. Building on this motivation, the specific research gap addressed in this study is the limited availability of rigorous, multi-country comparative evidence on whether hybrid SEIR–machine-learning models reliably outperform their mechanistic and purely data-driven counterparts under realistic, chronologically held-out forecasting conditions. The objectives of this study are therefore to: (i) develop and implement two complementary hybrid SEIR–machine-learning pathways (a machine-learned transmission rate and a machine-learned residual correction); (ii) evaluate these hybrid approaches against a mechanistic SEIR baseline, a purely data-driven Random Forest model, and a simple persistence baseline, using a consistent chronological train–test framework across four countries with heterogeneous epidemic trajectories; and (iii) identify the conditions under which hybrid AI components meaningfully improve, or alternatively degrade, forecasting performance relative to the mechanistic baseline. Correspondingly, we test the working hypothesis that integrating machine-learning components into a mechanistic SEIR framework improves short-term forecasting accuracy

and adaptability relative to the mechanistic baseline alone; as detailed in Sections 4 and 5, our results indicate that this hypothesis is not consistently supported by the data, and we discuss the implications of this finding for the design of future hybrid epidemic forecasting systems.

2. Related work

2.1. Mathematical models for infectious disease spread

Classical compartmental models such as SIR and SEIR describe the movement of individuals between susceptible, exposed, infectious and removed classes and remain the backbone of epidemic modeling [1]. During COVID-19, SEIR-type models were widely used to estimate the basic reproduction number, evaluate non-pharmaceutical interventions and analyze the role of undetected infections [2, 4, 14] (Gupta *et al.* [2] and Rai *et al.* [14] are preprints). Several extensions incorporated additional compartments for vaccination, hospitalization or environmental reservoirs [3, 15, 16].

These models offer clear epidemiological interpretation but usually rely on strong assumptions such as homogeneous mixing and piecewise-constant parameters. As highlighted by reviews on COVID-19 modeling, fixed-parameter SEIR models struggle to adapt to rapid changes in transmission caused by behavioral responses, policy changes and testing practices [5, 8].

2.2. Machine learning and deep learning for COVID-19

Many studies have explored machine-learning and deep learning methods for COVID-19 prediction. These methods are used for case forecasting and also for clinical applications. Time series models such as vector autoregressive (VAR) models, long short-term memory (LSTM), gated recurrent unit (GRU), eXtreme Gradient Boosting (XGBoost), and hybrid autoregressive integrated moving average–machine-learning (ARIMA–ML) approaches have been applied to predict daily cases or deaths in different countries [6, 7, 17–22]. Review studies show that tree-based models and deep neural networks can achieve strong short-term forecasting performance. At the same time these models can be sensitive to changes in data patterns and often provide limited interpretability [5, 8, 9]. Machine learning methods including Random Forest and structural equation modelling have been applied to identify COVID-19 risk and epidemiological factors [23].

Artificial intelligence techniques have also been used in medical imaging and diagnostic analysis, such as detecting COVID-19 from chest X-ray or CT images [24, 25]; this body of work is included only to illustrate the broader scope of AI applications in COVID-19 research and is not directly related to the country-level time-series forecasting methods used in this study.

2.3. Hybrid epidemic models

Several studies have proposed hybrid frameworks that combine mechanistic epidemic models with machine-learning methods. The goal is to connect interpretability with data-driven adaptability. In many studies, mechanistic models are improved using machine-learning-based corrections to reduce forecasting errors across different COVID-19 waves [10]. A hybrid framework called SIMLR integrates machine-learning within a classical SIR model to estimate time-varying transmission dynamics [11]. Other approaches combine mechanistic models with neural networks or boosted tree algorithms to capture residual patterns and include additional epidemiological signals [12, 13, 26, 27]. Recent review and position studies suggest that artificial intelligence can support traditional epidemic models by learning complex and evolving drivers such as mobility, climatic factors, and population behaviour. These approaches can improve forecasting performance while still maintaining transparency and interpretability [28, 29]. Following this hybrid modeling perspective, the present study adopts a simple and interpretable framework that combines a classical SEIR model with a tree-based residual learning approach for country-level COVID-19 forecasting.

3. Data and methods

3.1. Data source and study countries

This study uses publicly available country-level COVID-19 case data from the Johns Hopkins University CSSE repository [30]. The dataset provides daily cumulative confirmed cases for regions across the world. The analysis focuses on four countries that experienced different epidemic patterns and multiple waves during the period from 2020 to 2023. These countries are India, the United States, Italy, and Japan. These countries also provide long and continuous time series data that is suitable for both mechanistic modeling and machine learning based forecasting. For the SEIR model (Section 3.4), the total population N used for each country was: India, 1.38×10^9; the United States, 3.31×10^8; Italy, 6.0×10^7; and Japan, 1.26×10^8, based on the World Bank World Development Indicators, indicator SP.POP.TOTL ("Population, total"), 2021 estimates, accessed 5 July 2025 [31].

For each country, provincial cumulative case counts were combined to obtain national totals. Daily new confirmed cases were then calculated as first differences. A 7-day moving average was applied to reduce noise in the reported data, which can arise from weekend reporting delays, retrospective corrections, and other administrative variations [5, 8]. Days that reported negative values were set to zero because these values usually occur after retrospective adjustments in the data. The study period ranges from 22 January 2020 to 9 March 2023, comprising 1,143 daily observations. This start date corresponds to the first date for which the Johns Hopkins University CSSE repository records confirmed COVID-19 case data across all four study countries. Because every machine-learning and hybrid model in this study constructs lagged predictors up to 28 days and discards the resulting missing rows, the usable modeling sample for these components begins 28 days later, on 19 February 2020, and spans 1,115 days. This distinction is made explicit here to reconcile the raw data period with the train/test split sizes reported in Section 3.5.

To illustrate epidemic progression Figure 1 (a) shows the unsmoothed daily incidence curves for the four countries.

Table 1: Notation and parameter definitions used in the proposed epidemic forecasting framework.

Symbol	Description
$S(t)$	Number of susceptible individuals at time t
$E(t)$	Number of exposed (infected but not yet infectious) individuals at time t
$I(t)$	Number of infectious individuals at time t
$R(t)$	Number of removed (recovered or deceased) individuals at time t
N	Total population size
$\beta(t)$	Time-varying transmission rate
$\widehat{\beta}(t)$	Machine-learned estimate of the transmission rate
σ	Incubation rate ($\sigma = 1/4.5$)
γ	Recovery rate ($\gamma = 1/6$)
$R_t(t)$	Effective reproduction number at time t
ρ	Reporting fraction linking true infections to observed cases
ℓ	Reporting delay (in days)
$y_{\text{obs}}(t)$	Observed daily confirmed COVID-19 cases at time t
$y_{\text{SEIR}}(t)$	Incidence implied by the SEIR model
$r(t)$	Residual error between observed and SEIR-implied incidence
$\widehat{y}(t)$	Final hybrid model prediction
MAE	Mean absolute error
RMSE	Root mean squared error
MAPE	Mean absolute percentage error

These curves highlight the timing and intensity of major waves. The plots show the strong Delta wave in India, several sustained peaks in the United States, repeated seasonal waves in Italy and a prolonged Omicron driven surge in Japan. Figure 1 (b) presents cumulative case counts over the same period. This figure helps show long-term growth patterns and differences in total disease burden across countries.

3.2. Model notation and parameter definitions

Table 1 lists the key symbols and parameters used in the mathematical hybrid and machine-learning parts of this study.

3.3. Data preprocessing and observation model

The accuracy of epidemic modeling depends on the availability of clean and reliable input data. However, the raw COVID-19 case counts reported in the Johns Hopkins CSSE dataset often exhibit reporting irregularities, including reduced case counts during weekends, sudden increases caused by backlog updates and occasional negative values resulting from retrospective data corrections. These inconsistencies have been widely documented and can adversely affect the reliability and predictive performance of both epidemiological and machine-learning models [5, 28].

To ensure that our analysis is based on a stable and meaningful signal, we apply a structured preprocessing pipeline that removes these inconsistencies and produces a smooth daily case series suitable for SEIR modeling and for use as predictors in machine-learning models.

3.3.1. Extraction of daily cases

Daily incidence was calculated from cumulative confirmed cases $C(t)$ using a simple finite difference approach with a non-negativity constraint:

$$\text{DailyRaw}(t) = \max\{0,\ C(t) - C(t-1)\}. \tag{1}$$

The reported data may contain negative values because of retrospective corrections. These values were therefore replaced with zero. Figure 2 shows the effect of this preprocessing procedure for the four study countries. The upper panels present cumulative confirmed cases while the lower panels show the corresponding daily incidence after applying a seven-day moving average smoothing method. This visualization shows how smoothing reduces reporting noise and reveals the underlying epidemic trend which is later used for SEIR calibration and machine-learning feature construction.

3.3.2. Winsorization and smoothing

Daily incidence values occasionally include extreme outliers caused by bulk reporting. To reduce their influence without discarding genuine information, the series was winsorized at the 99th percentile:

$$\text{DailyWinz}(t) = \min\{\text{DailyRaw}(t),\ Q_{0.99}\}. \tag{2}$$

This threshold affects only the most extreme 1% of days in each country's series (12 days out of 1143 total days, 1.05%, identical across all four countries). The 99th-percentile cutoff was chosen to dampen isolated bulk-reporting artifacts (e.g., single-day backlogs) while preserving genuine epidemic peaks, since true wave peaks in COVID-19 incidence are typically sustained over multiple days rather than concentrated in a single extreme observation. We note that with only 12 affected days per country, and a relatively close gap between the 99th-percentile threshold and the unwinsorized maximum (e.g., India: $Q_{0.99}$ = 361,971 vs. maximum= 414,188), the winsorization has limited practical effect, and we would not expect substantively different results without it; a full re-fit with and without winsorization is identified as a direction for future sensitivity analysis. To further assess the sensitivity of our preprocessing choices, we examined how peak timing and peak magnitude in the smoothed incidence series respond to alternative winsorization thresholds (95th, 99th percentile, and no winsorization) combined with alternative smoothing windows (5-day, 7-day, and 14-day centered moving averages), for all four countries. Across all nine parameter combinations tested per country, the identified epidemic peak consistently fell within the same known pandemic wave (e.g., India's peak was always identified within the April-May 2021 Delta wave window, and Japan's within the July–August 2022 wave window), indicating that peak *timing* is robust to these preprocessing choices. Smoothing window choice (5 vs. 7 vs. 14 days) had a negligible effect on the mean signal level (typically under 0.1% difference). winsorization threshold had a more pronounced effect on

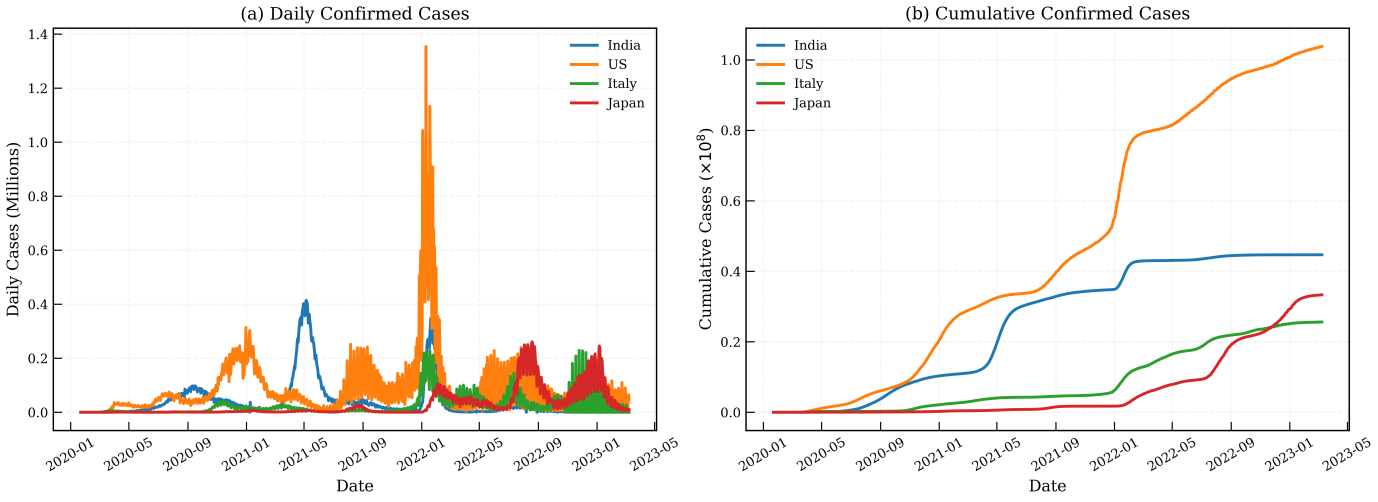

Figure 1: Daily and cumulative confirmed COVID-19 cases for India, the United States, Italy, and Japan.

peak *magnitude*: the 95th-percentile threshold suppressed peak values substantially (e.g., by approximately 45% for India relative to no winsorization), whereas the 99th-percentile threshold used in this study produced peak magnitudes much closer to the unwinsorized series (e.g., within approximately 8% for India), supporting the 99th percentile as a reasonable compromise between outlier suppression and peak preservation. The complete re-fitting of the SEIR model under each preprocessing configuration remains an important aspect for future sensitivity analysis:

$$\text{DailySmoothed}(t) = \frac{1}{7}\sum_{k=-3}^{3} \text{DailyWinz}(t+k). \tag{3}$$

The smoothed curve serves as the primary observational signal used in all subsequent modeling steps.

3.3.3. Observation model

To link the observed incidence to latent epidemic dynamics, we relate new reported cases to the outflow from the exposed compartment of an SEIR model. Let $\sigma = 1/4.5$ be the mean incubation rate, ℓ denote the reporting delay, and ρ denote the country-specific reporting fraction (formally estimated in Section 3.4). The observation model is:

$$y_{\text{obs}}(t) \approx \rho\,\sigma\,E(t-\ell). \tag{4}$$

Thus a direct empirical proxy for the shifted exposed compartment is:

$$\widehat{E}(t) = \frac{y_{\text{obs}}(t)}{\rho\,\sigma}. \tag{5}$$

The delay parameter ℓ was estimated separately for each country by selecting, from a candidate set $\ell \in \{3,4,5,6,7\}$, the value that minimizes the sum of squared error between the observed incidence $y_{\text{obs}}(t)$ and the model-implied incidence $\rho\,\sigma\,E(t-\ell)$ over the training window, as detailed in Section 3.4.

3.3.4. Correlation structure and feature justification

To investigate the day-to-day evolution of case numbers, we analyzed the correlation between the raw daily series, the smoothed 7-day series, and their lagged versions at 1, 7, and 14 days. All four countries showed a strong correlation at a lag of 7 days, which suggests a clear weekly reporting pattern in the surveillance data. This weekly cycle has been widely observed in COVID-19 reporting and helps explain why lag 7 and lag 14 features work well as predictors in the machine-learning and hybrid models used in this study. To illustrate these patterns across all study countries, the correlation heatmaps for India, the United States, Italy, and Japan are presented in Figure 3.

3.4. Baseline SEIR modeling framework

Mechanistic compartmental models provide a structured and interpretable representation of infectious disease transmission, and they have been widely applied in COVID-19 forecasting [1, 5, 28]. To establish a mechanistic baseline against which the hybrid and machine-learning approaches can be compared, we fit a discrete-time SEIR model with a time-varying transmission rate $\beta(t)$ to the preprocessed incidence series obtained in Section 3.3.

3.4.1. Model structure

The SEIR system divides the population into four compartments: susceptible (S), exposed (E), infectious (I), and removed (R). The continuous-time dynamics are described by:

$$\frac{dS}{dt} = -\beta(t)\frac{SI}{N}, \qquad \frac{dE}{dt} = \beta(t)\frac{SI}{N} - \sigma E, \tag{6}$$

$$\frac{dI}{dt} = \sigma E - \gamma I, \qquad \frac{dR}{dt} = \gamma I. \tag{7}$$

For all four countries, the simulation is initialized at the start of the study period (22 January 2020) with $E(0) = 1$, $I(0) = 1$, $R(0) = 0$, and $S(0) = N - E(0) - I(0) - R(0)$, i.e., a single exposed and a single infectious individual at the outset, with

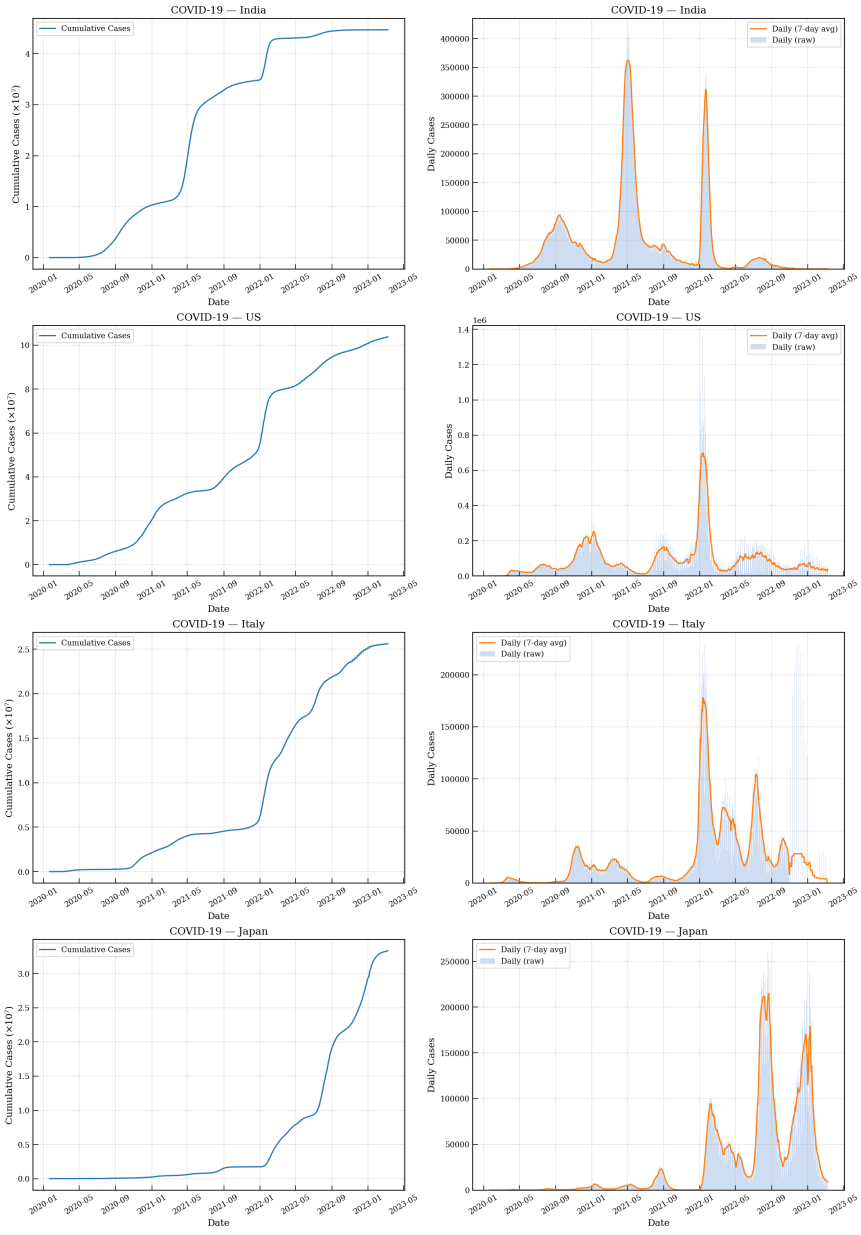

Figure 2: Country-level COVID-19 trends showing cumulative confirmed cases (top row) and daily new cases with 7-day moving average smoothing (bottom row) for India, the United States, Italy, and Japan.

the remaining population susceptible. These initial conditions are held fixed across countries and are not separately estimated.

Following standard COVID-19 modeling studies [4, 8], we adopt fixed epidemiological parameters:

$$\sigma = 1/4.5, \qquad \gamma = 1/6.$$

These values correspond to a mean incubation period of 4.5 days and a mean infectious period of 6 days, which fall within the range commonly reported across the COVID-19 modeling literature for the ancestral and early variant periods (incubation estimates of approximately 4–6 days, infectious/recovery periods of approximately 5–10 days), and are adopted here as fixed, literature-consistent averages for tractability across all four countries and the full multi-wave study period. We acknowledge that biological parameters such as incubation and infectious periods are known to have varied across SARS-CoV-2 variants (e.g., Alpha, Delta, Omicron) and may also have been influenced by changing isolation policies and reporting practices; holding σ and γ fixed is therefore a simplifying assumption, and wave-specific or time-varying recalibration of these parameters is identified as a direction for future sensitivity analysis.

For numerical stability, a discrete-time version with daily updates is implemented.

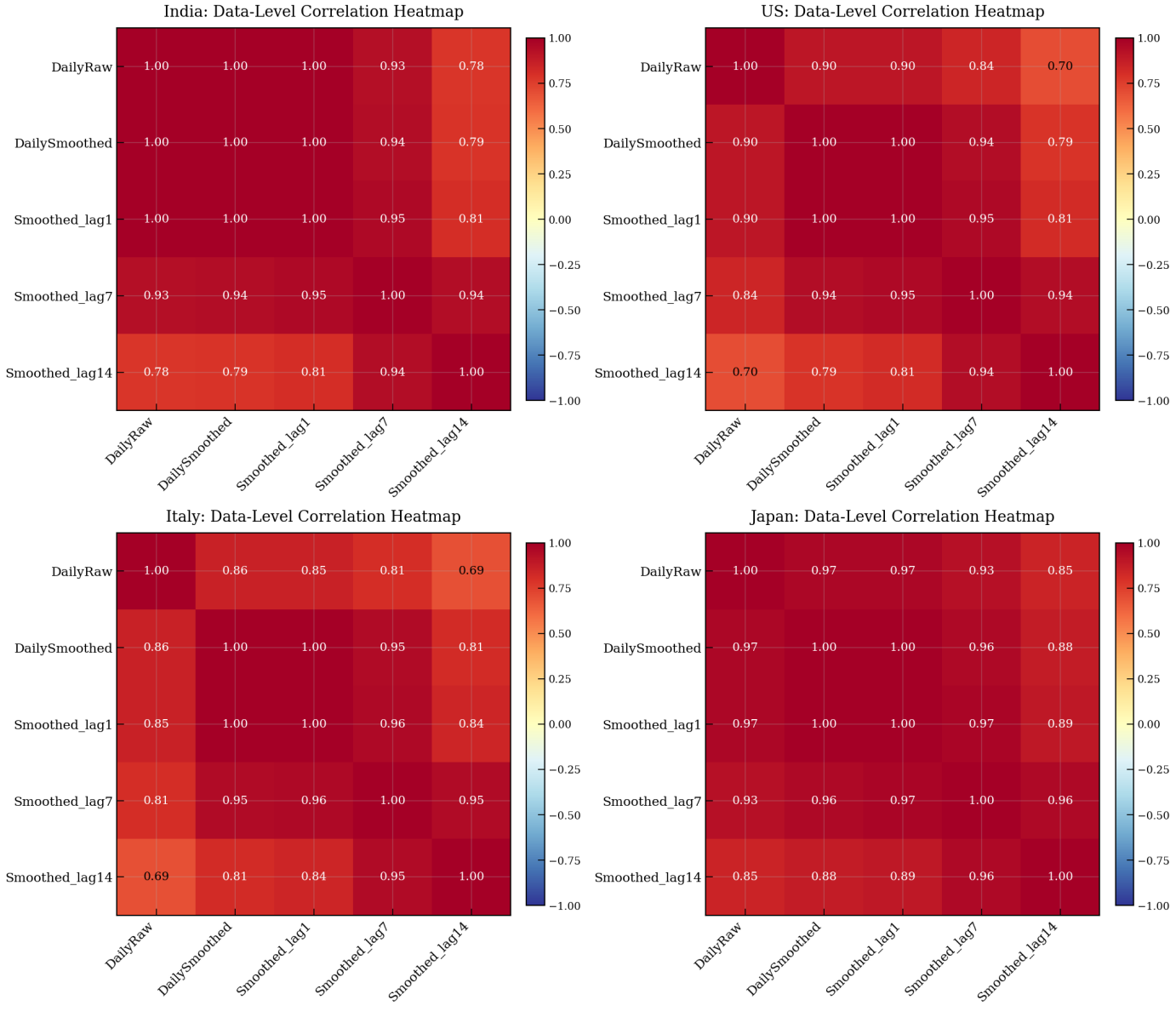

Figure 3: Correlation heatmaps for the five incidence-derived features (raw daily cases, smoothed series, and 1-day, 7-day, and 14-day lagged smoothed values) across the four study countries.

3.4.2. *Observation model and reporting delay*

Reported cases do not directly correspond to new infections; instead, they reflect delayed and partially observed incidence. To link the SEIR model to the smoothed observational series, we use a standard delayed-incidence observation model [28]:

$$y_{\text{obs}}(t) \approx \rho\, \sigma\, E(t - \ell), \tag{8}$$

where ρ represents the reporting fraction and ℓ is a reporting delay measured in days. Both parameters are country-specific. We evaluate delays $\ell \in \{3, 4, 5, 6, 7\}$ and select the value minimizing the fitted error.

3.4.3. *Time-varying transmission rate*

To capture changes in transmission caused by policy interventions, behavioral shifts, and the emergence of new variants, $\beta(t)$ is modelled as a cubic spline in log-space with knots spaced at 14-day intervals:

$$\log \beta(t) = \text{Spline}_\theta(t), \tag{9}$$

a smooth and flexible parameterization commonly used in real-time epidemic inference [26]. With knots placed every 14 days across the training window only (22 January 2020 to 29 July 2022, 920 days), this yields 67 spline parameters θ per country. The spline coefficients are fitted by minimizing the sum of squared error between the model-implied incidence and the observed incidence (Equation 8), restricted strictly to

the training window, using a standard limited-memory quasi-Newton optimization method with box constraints, with convergence assessed by a maximum of 300 iterations and a function-value tolerance of 10^{-9} between successive iterations. To discourage overfitting and excessively sharp changes in transmission between adjacent knots, a second-difference smoothing penalty is added to the objective function, $\lambda \sum_k (\Delta^2 \theta_k)^2$ with $\lambda = 0.01$, which penalizes curvature in $\log \beta(t)$. For the 223-day held-out test window (30 July 2022 to 9 March 2023), $\beta(t)$ is held fixed at its last training-fitted value (flat carry-forward), so that no test-period information enters the fit at any point. The effective reproduction number is computed as:

$$R_t(t) = \frac{\beta(t)}{\gamma}.$$

3.4.4. Country-level fitted parameters

The optimal delay ℓ and estimated reporting fraction ρ for each country are:

The fitted reporting delays and reporting fractions were India, $\ell = 7$ days and $\hat{\rho} = 0.040$; the United States, $\ell = 6$ days and $\hat{\rho} = 0.416$; Italy, $\ell = 4$ days and $\hat{\rho} = 0.492$; and Japan, $\ell = 5$ days and $\hat{\rho} = 0.153$.

These fitted values of ρ are model-dependent scaling parameters and should be interpreted cautiously rather than as direct measurements of reporting completeness. While the comparatively low value estimated for India ($\hat{\rho} = 0.040$) is broadly consistent with previously reported under-ascertainment in large, high-burden transmission settings, and the higher values for the United States and Italy are broadly consistent with more complete surveillance systems, confirming the actual magnitude of under-reporting in any of these countries would require external validation against independent data sources such as seroprevalence surveys, which is beyond the scope of the present study.

3.4.5. Fitted incidence and reproduction number

The fitted SEIR model provides a mechanistic reconstruction of the epidemic trajectories in all four study countries. Figure 4 presents the comparison between the observed smoothed incidence and the SEIR-implied incidence, whereas Figure 5 displays the corresponding time-varying reproduction number $R_t(t)$ estimated across the full study period.

Overall, the baseline SEIR model is able to reproduce the timing and relative magnitude of the major COVID-19 waves, despite the presence of substantial reporting variability in the raw data. The fitted model therefore serves as a transparent and interpretable benchmark against which the performance of the hybrid and machine-learning forecasting components can be evaluated in later sections.

The time-varying transmission rate $\beta(t)$ for the SEIR baseline is estimated using data strictly from the training window (22 January 2020 – 29 July 2022), with the value held fixed at its last training-fitted level throughout the 223-day held-out test window (30 July 2022 – 9 March 2023). The SEIR baseline reported in this study therefore represents a genuine out-of-sample forecast over the test window, consistent with the persistence, Random Forest, Track A, and Track B evaluations reported in Section 3.5.3. As detailed in Section 4, this train-only fitting procedure results in substantially weaker test-window performance than would be obtained from a full-period (leakage-containing) fit, since a transmission rate frozen at its late-training value cannot anticipate subsequent epidemic waves. The baseline SEIR model successfully captures the timing and general magnitude of major COVID-19 waves and serves as a mechanistic benchmark for evaluating hybrid and machine-learning forecasting approaches.

3.5. Hybrid SEIR–machine-learning framework

The purpose of introducing machine-learning components is not to replace the SEIR model. The aim is to examine whether artificial intelligence can improve predictive flexibility during periods of rapid behavioral change, reporting irregularities, or intervention-driven shifts. This hybrid design aligns with the broader objective of AI-based predictive modeling by identifying when and how data-driven learning can meaningfully complement epidemiological structure.

To bridge the gap between mechanistic structure and data-driven flexibility, we develop a hybrid modeling framework that integrates outputs from the fitted SEIR model with machine learning (ML) components. The guiding principle is to preserve the interpretability and epidemiological grounding of SEIR dynamics, while enabling the model to adapt to fine-scale patterns in the data that mechanistic models alone cannot capture.

Two complementary hybrid pathways are examined:

Track A learns a data-driven transmission rate $\beta(t)$ using lagged epidemiological features and reconstructs incidence via SEIR simulation. Track B models the residual error between SEIR-implied incidence and the smoothed observational series.

All five models are evaluated on a single, common chronological test window of 30 July 2022 to 9 March 2023 (223 days), reconciling the 1,115-day usable modeling sample described in Section 3.3. For Track A and Track B, the training window spans the full pre-test raw series, 22 January 2020 to 29 July 2022 (920 days); since neither component requires the 28-day lag-feature burn-in, no observations are discarded prior to training. For the pure machine-learning (Random Forest) model in Section 3.5.4, training and validation are drawn from the 1,115-day usable sample (19 February 2020 onward, after the 28-day lag burn-in) preceding the common test window, split approximately 75%/25% into training and validation subsets. In all cases, hyperparameters (e.g., Random Forest tree depth and number of estimators) were fixed prior to model fitting and were not tuned using the test window; the validation subset for the Random Forest model is used exclusively to estimate bootstrap residuals for prediction intervals and is kept strictly separate from the held-out test window used for performance evaluation.

3.5.1. Track A: Learning a data-driven transmission rate

In Track A, we replace the spline-based transmission model from Section 3.4 with a machine-learning estimator that predicts $\log \beta(t)$ using lagged values of the observed incidence and

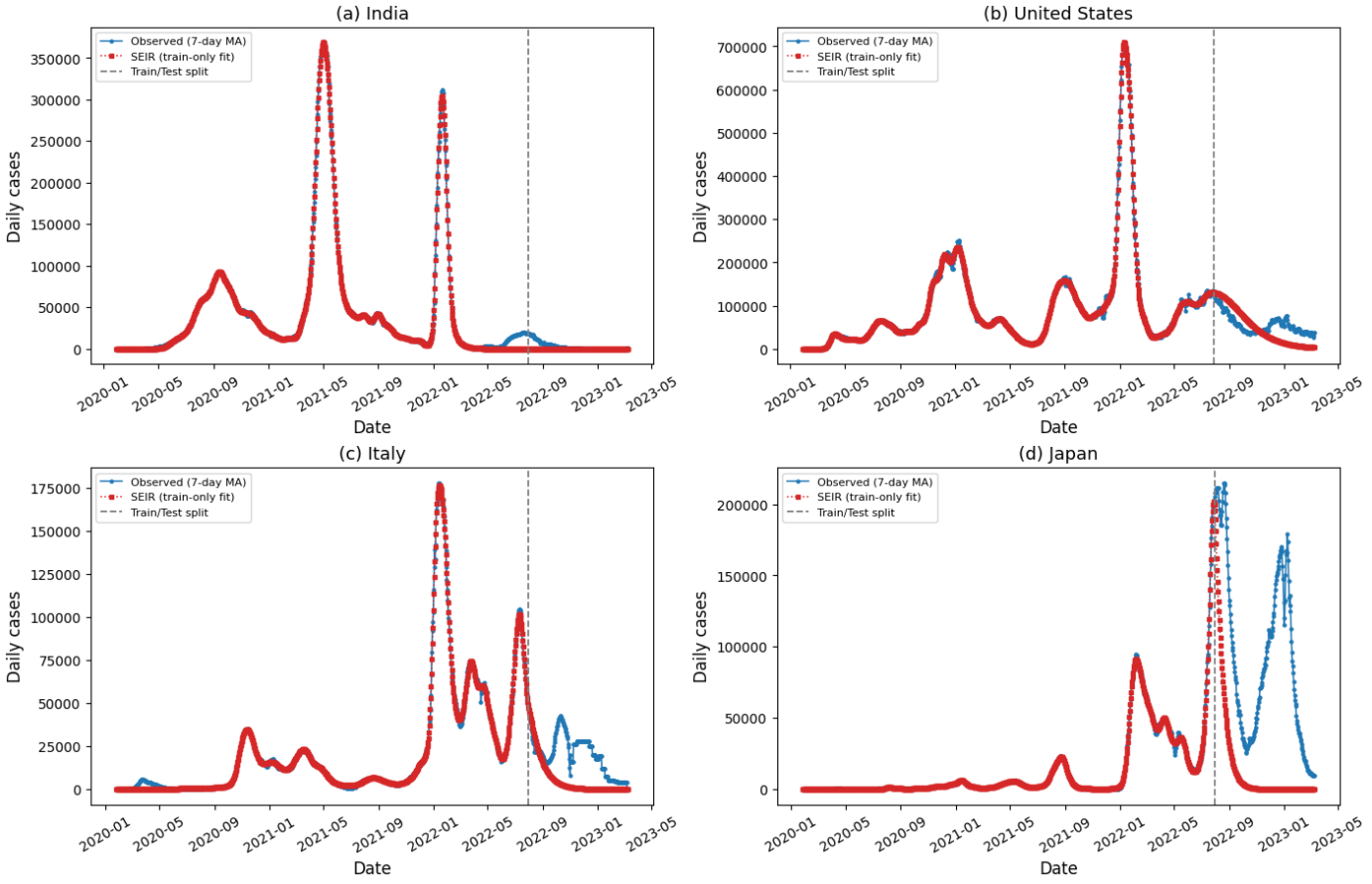

Figure 4: Observed smoothed incidence (7-day moving average) compared with SEIR-implied incidence for India, the United States, Italy, and Japan. The transmission rate $\beta(t)$ is fitted using training data only (22 January 2020 – 29 July 2022) and held fixed at its last-fitted value thereafter; the vertical dashed line marks the train/test split (30 July 2022). The test-window portion (right of the dashed line) represents a genuine out-of-sample forecast, evaluated formally in Table 2.

the reproduction number $R_t(t)$. The estimated $\widehat{\beta}(t)$ is subsequently passed through the SEIR system, yielding a fully mechanistic reconstruction driven by a data-informed transmission surface. The $R_t(t)$ feature used as a Track A predictor is derived from the SEIR baseline's $\beta(t)$ spline (Section 3.4), which is fitted using training data only (before 30 July 2022); accordingly, the $R_t(t)$ values used as predictors during the test window contain no future information, and the Track A evaluation reported here represents a genuine out-of-sample forecast, consistent with the other models compared in Table 2.

A country-specific reporting scale is then estimated by least-squares fitting between the SEIR implied incidence and the smoothed observational series, restricted to the training window. This approach preserves the epidemiological structure of the model while allowing transmission dynamics to reflect subtle, rapid changes captured in the data.

Figure 6 presents the Track A reconstructions for all four countries, illustrating the close correspondence between observed and hybrid-fitted incidence during major epidemic waves. Although Track A does not consistently reduce forecast error, it demonstrates how AI-driven transmission learning can replicate epidemic timing without violating epidemiological constraints, thereby offering interpretability-preserving adaptability rather than purely error-driven optimization

3.5.2. *Track B: Residual learning to correct SEIR mismatch*

Residual learning provides an alternative way of combining mechanistic and data-driven components. Let

$$r(t) = y_{\text{obs}}(t) - y_{\text{SEIR}}(t),$$

denote the discrepancy between observed and SEIR-implied incidence. Track B trains a machine-learning regressor to learn $r(t)$ using lagged incidence and lagged R_t features, after which the hybrid prediction is obtained as

$$\widehat{y}(t) = y_{\text{SEIR}}(t) + \widehat{r}(t).$$

As shown in Section 4, residual learning produces a small but consistent RMSE improvement over the SEIR baseline in all four countries once evaluated on the common held-out test window. This modest, consistent gain suggests that a residual correction term can partially offset the mechanistic model's inability to track new waves, though the improvement remains

Table 2: Complete comparison of forecasting performance across models and countries on the common held-out test window (30 July 2022 – 9 March 2023, 223 days). MAE and RMSE are in daily case counts; MAPE and high-incidence MAPE (computed only on days with observed incidence ≥ 100 cases) are in percent; peak-timing error is in days. Random Forest predictions are clipped at zero (non-negative) before metric computation.

Country	Model	MAE	RMSE	R^2	MAPE (%)	High-incidence MAPE (%)	Peak-timing error (days)
India	SEIR (baseline)	3121.62	5699.04	−0.429	100.00	100.00	0
	Track A (ML $\beta(t)$)	3121.62	5699.04	−0.429	100.00	100.00	0
	Track B (Residual)	4377.10	4957.53	−0.081	1021.34	973.19	2
	Persistence baseline	178.71	430.10	–	6.81	6.85	0
	Pure ML (Random Forest)	5875.37	6142.22	–	2010.63	1904.34	0
United States	SEIR (baseline)	29660.28	32580.98	−1.173	57.11	57.11	0
	Track A (ML $\beta(t)$)	29754.43	32670.20	−1.185	57.31	57.31	0
	Track B (Residual)	28303.35	31252.23	−0.999	54.96	54.96	0
	Persistence baseline	2206.92	3328.10	–	4.12	4.12	0
	Pure ML (Random Forest)	30247.33	34354.76	–	57.10	57.10	1
Italy	SEIR (baseline)	15231.85	19065.10	−1.743	78.89	78.89	0
	Track A (ML $\beta(t)$)	15255.81	19067.45	−1.744	79.03	79.03	0
	Track B (Residual)	13874.61	18532.09	−1.592	56.83	56.83	0
	Persistence baseline	758.70	1732.54	–	3.90	3.90	0
	Pure ML (Random Forest)	7174.20	8479.67	–	67.29	67.29	120
Japan	SEIR (baseline)	76392.61	90191.22	−1.107	89.68	89.68	22
	Track A (ML $\beta(t)$)	76192.67	90026.60	−1.099	89.38	89.38	22
	Track B (Residual)	75809.79	89434.62	−1.072	89.33	89.33	22
	Persistence baseline	3301.80	4752.78	–	4.01	4.01	1
	Pure ML (Random Forest)	77270.35	98501.86	–	71.39	71.39	184

far smaller than the gap to the persistence baseline. Figure 7 illustrates the residual-corrected trajectories across all four countries.

3.5.3. Evaluation on hold-out test windows

To assess predictive skill, all baseline, hybrid, and machine-learning configurations are evaluated on the final 20% of the time series for each country using a strict chronological train–test split. Performance is assessed using multiple complementary metrics, including the mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), peak-timing error, and MAPE evaluated during high-incidence periods (days with observed incidence ≥ 100 cases). High-incidence MAPE is computed as $\frac{1}{n}\sum|y_{\text{obs}}(t)-\hat{y}(t)|/y_{\text{obs}}(t)\times 100$, using the observed incidence $y_{\text{obs}}(t)$ as the denominator, restricted to test-window days for which $y_{\text{obs}}(t) \geq 100$. Of the 223 test-window days per country, 218 (97.8%) qualified for India, 223 (100.0%) for the United States, 220 (98.7%) for Italy, and 223 (100.0%) for Japan. Because nearly all test-window days qualify as high-incidence for every country in this study, the high-incidence restriction has only a minor effect on the number of days retained, and cross-country comparability of this metric is not meaningfully affected by differing numbers of low-incidence days. This conditional evaluation of percentage-based error nonetheless addresses the well-known issue of inflated MAPE values under very small denominators.

We note that all forecasting metrics reported in this study, including those in Table 2, are computed against the smoothed (7-day moving average) incidence series $y_{\text{obs}}(t)$ described in Section 3.3, rather than raw daily case counts. This should be considered when comparing the error magnitudes reported here with studies that evaluate forecasts directly against raw reported cases, which typically exhibit substantially higher day-to-day variance; smoothing tends to reduce both the apparent volatility of the target series and, correspondingly, the absolute magnitude of forecasting errors relative to what would be observed on raw incidence.

The comparative behaviour of all models over the hold-out

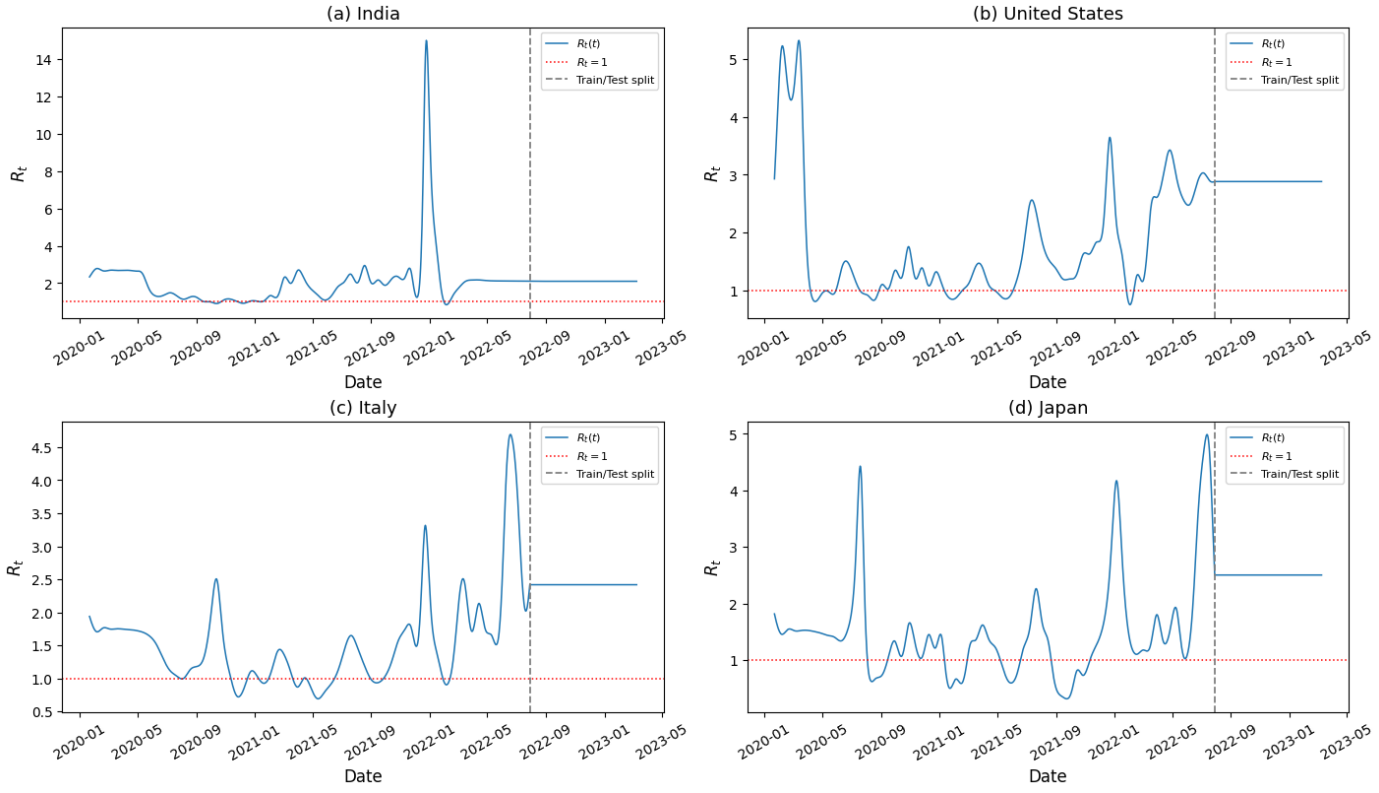

Figure 5: Estimated time-varying reproduction number (R_t) for India, the United States, Italy, and Japan obtained from the fitted SEIR model, fitted on training data only. The vertical dashed line marks the train/test split (30 July 2022); $R_t(t)$ is held constant at its last-fitted value for the test window. The red dashed horizontal line represents the epidemic threshold ($R_t = 1$).

test window is examined qualitatively using the test-window forecasting results (Figure 8) and the short-term projection experiments. These figures show the differences in stability, peak alignment, and uncertainty between the baseline SEIR model, the two hybrid extensions, and the purely data-driven forecasting approach.

The baseline SEIR model shows the most stable behavior during periods with high case incidence. It maintains consistent peak timing and smooth epidemic trends. Track A provides modest improvements in short-term flexibility by allowing data-driven adjustment of the transmission rate. These improvements are not consistent across all countries. Track B follows the SEIR trajectory closely and produces a small, consistent RMSE improvement over SEIR in every country, though the residual correction is too small to bring its performance close to the persistence baseline.

3.5.4. Machine-learning forecasting and 30-day projections

The persistence baseline used throughout this study is a one-step- ahead forecast, defined as $\hat{y}(t) = y_{\text{obs}}(t-1)$: the true previously observed incidence value (not the model's own prior prediction) is used as the forecast for the following day. The forecast horizon is one day, re-initialized with the true observation at every step across the full 223-day test window; it is not a fixed multi-step forecast.

Beyond hybrid modeling, we also evaluate a pure machine-learning forecasting model (Random Forest) trained on lagged incidence, lagged reproduction numbers, and simple temporal features. This model produces one-step-ahead predictions over the test window and a 30-day-ahead probabilistic forecast generated via bootstrap resampling of validation residuals. Prediction intervals for both the test-window evaluation and the 30-day forecast are constructed via residual bootstrap resampling. Residuals are first computed on the validation set as the difference between observed incidence and the model's validation-set predictions, $y_{\text{val}}(t) - \hat{y}_{\text{val}}(t)$. For each of 500 bootstrap iterations, residuals are sampled with replacement from this validation-residual pool and added to the point prediction at each time step; the resulting 500 simulated trajectories are used to construct a 90% prediction interval by taking the 5th and 95th percentiles at each time point. Point predictions and prediction-interval bounds are clipped at zero to ensure epidemiologically meaningful, non-negative incidence forecasts.

For reproducibility, the Random Forest model was implemented using scikit-learn (version 1.6.1, with NumPy 2.0.2 and pandas 2.2.2), with hyperparameters fixed in advance at 400 trees (`n_estimators=400`) and a maximum depth of 8 (`max_depth=8`); these values were not tuned via a systematic search procedure. A fixed random seed (`random_state=42`, and `np.random.seed(42)` / `random.seed(42)` for the bootstrap resampling) was used throughout to ensure reproducible results.

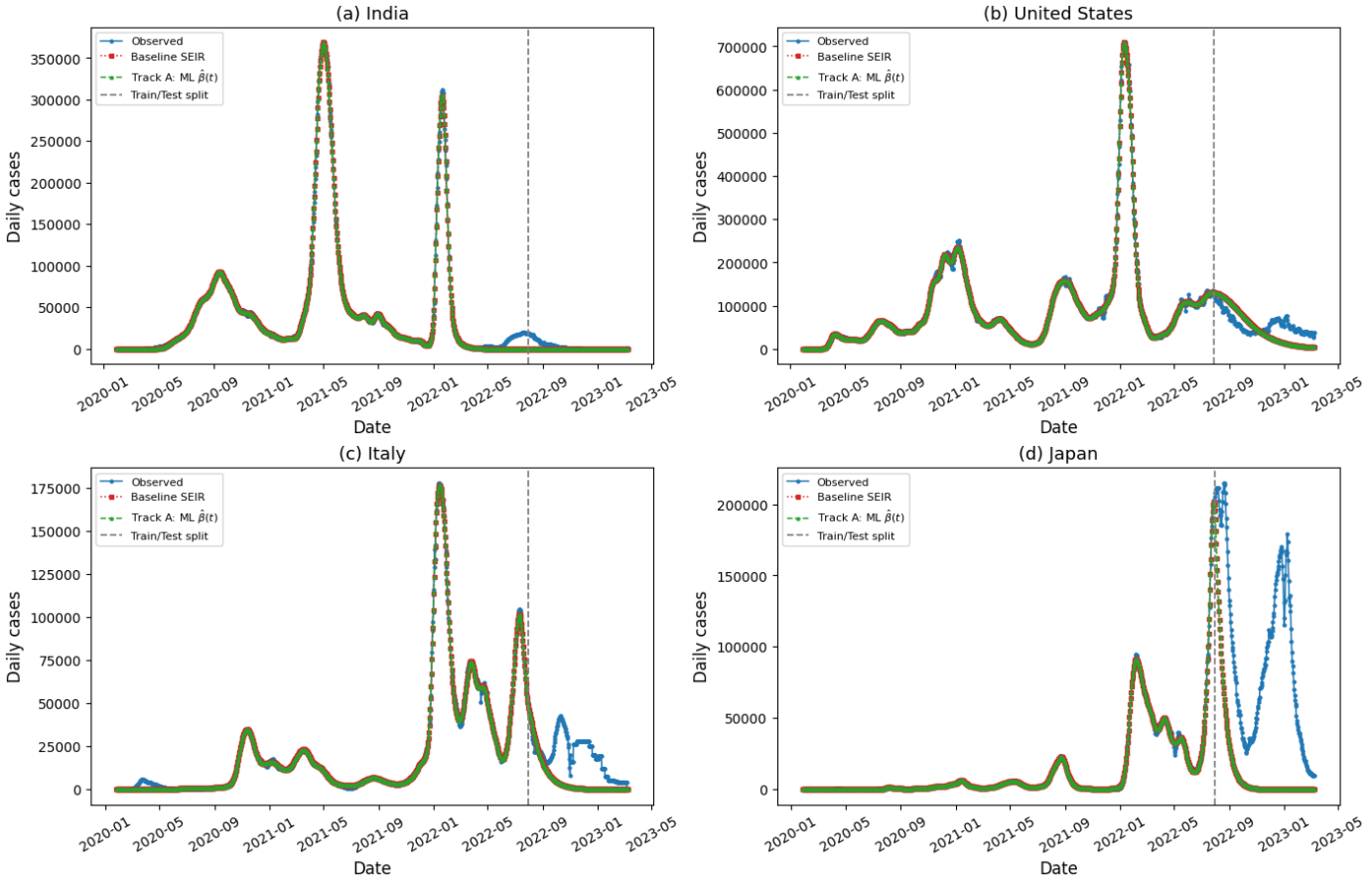

Figure 6: Hybrid Track A results showing SEIR reconstructions using the machine-learned transmission rate $\widehat{\beta}(t)$ for India, the United States, Italy, and Japan, trained strictly on data before 30 July 2022 (vertical dashed line). The test-window portion represents a genuine out-of-sample forecast, evaluated formally in Table 2.

The model was trained on an 11-dimensional feature set comprising lagged observed incidence and lagged effective reproduction number at four lags each (7, 14, 21, and 28 days: $y_{\text{lag\_7}}$, $y_{\text{lag\_14}}$, $y_{\text{lag\_21}}$, $y_{\text{lag\_28}}$, $R_{t,\text{lag\_7}}$, $R_{t,\text{lag\_14}}$, $R_{t,\text{lag\_21}}$, $R_{t,\text{lag\_28}}$), together with day-of-week, calendar month, and a linear time-trend index, using the approximately 75%/25% chronological train–validation split described in Section 3.5, drawn from the 1,115-day usable sample preceding the common 223-day test window. These features were selected a priori based on domain knowledge and the correlation analysis presented in Section 3.3.4, which identified 7-day and 14-day lags as showing the strongest correlation with current incidence across all four countries; no automated feature-selection algorithm (e.g., recursive feature elimination or importance-based filtering) was applied, and all 11 features were retained for model training.

The 30-day forecast begins the day immediately following the end of the study period (9 March 2023) and extends through 8 April 2023, identically for all four countries. This forecast does not use a strict one-step-at-a-time recursive procedure in which each day's model prediction feeds into the next day's lag features; instead, lagged incidence and R_t features for the 30-day horizon are constructed using a persistence-based proxy (the last observed value carried forward) rather than the model's own predictions or any observed future values, after which the model generates predictions for all 30 days in a single batch call. No observed future incidence values are used as inputs at any point in this procedure. We acknowledge that a strict recursive forecasting scheme, in which each predicted value updates the lag features used for the subsequent day, represents a more standard approach and is identified as a refinement for future work.

Figures 8 and 9 display, respectively, the test-window tracking performance and short-term forecasts for each country, including 90% prediction intervals. The pure ML model shows substantially reduced reliability across all four study countries during the test window, with MAPE values ranging from 57.1% (US) to 2010.6% (India) on the common test window (Table 2). The Random Forest RMSE exceeds the persistence baseline's RMSE in all four countries, but the size of this gap varies substantially: approximately 14x for India, 10x for the United States, 5x for Italy, and 21x for Japan, particularly during the rapid transitions and abrupt epidemic peaks visible for Japan and Italy in Figure 8. These results reinforce the value of mechanistic structure for medium-range epidemic prediction. As described in Section 3.5, all models (SEIR, Track A, Track B, the persistence baseline, and Random Forest) are evaluated over the

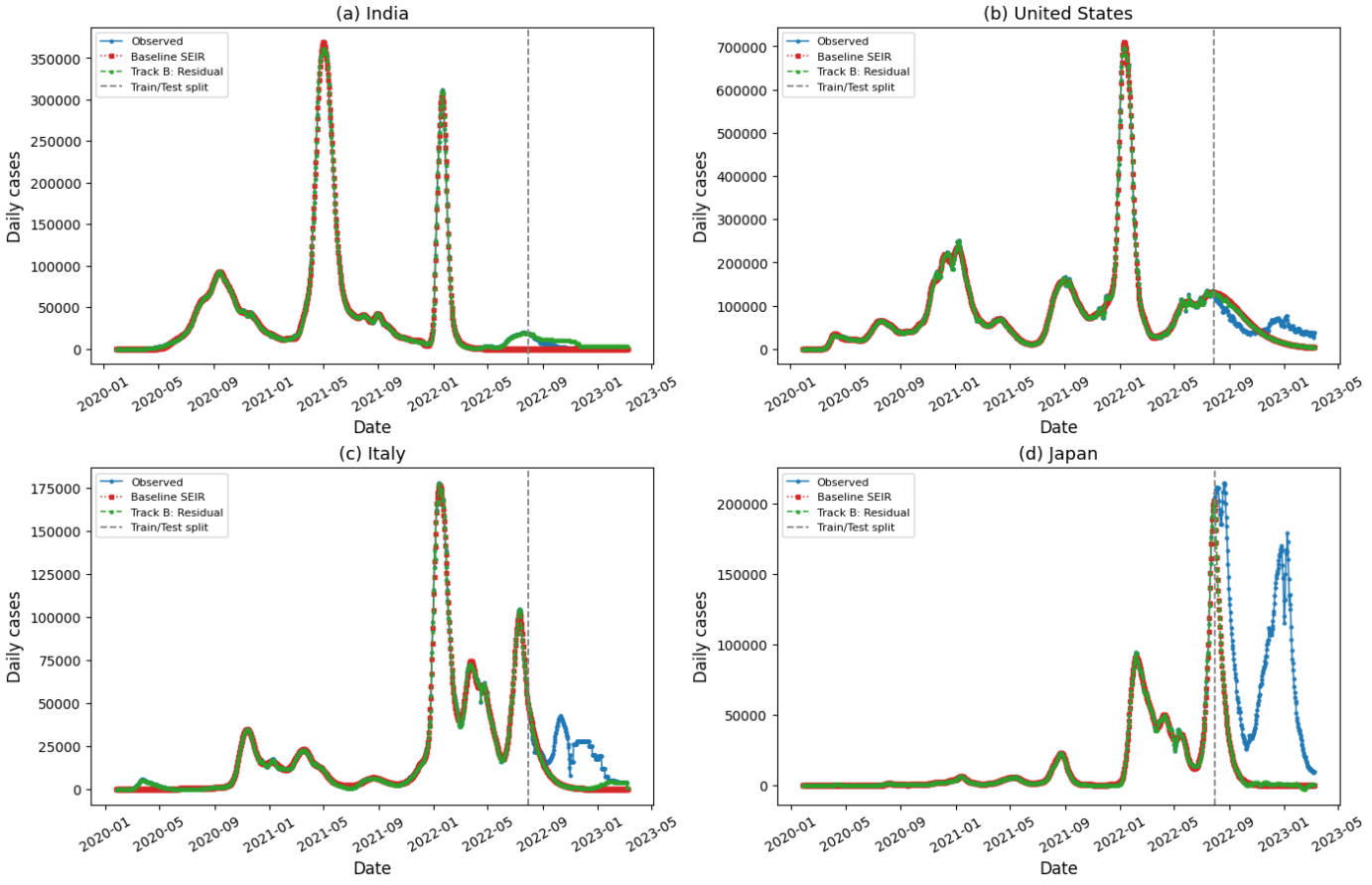

Figure 7: Hybrid Track B results showing residual-corrected SEIR incidence trajectories ($\hat{y}(t) = y_{SEIR}(t) + \hat{r}(t)$) for India, the United States, Italy, and Japan, with the residual model trained strictly on data before 30 July 2022 (vertical dashed line). The test-window portion represents a genuine out-of-sample forecast, evaluated formally in Table 2.

identical common test window of 30 July 2022 to 9 March 2023 (223 days) for the purposes of Table 2 and the Diebold–Mariano tests.

4. Results and discussion

The forecasting performance of the baseline SEIR model, the hybrid SEIR-AI extensions, and the purely data-driven model was evaluated using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). These metrics are commonly used in comparative epidemic forecasting studies. Across the four countries, the baseline SEIR model showed the most stable and consistent results. The RMSE values on the common 223-day held-out test window were 5699.04 for India, 32580.98 for the United States, 19065.10 for Italy, and 90191.22 for Japan.

These results suggest that the mechanistic SEIR framework can capture the timing and relative magnitude of major epidemic waves while avoiding large prediction deviations. Estimated reporting fractions also showed noticeable differences across countries. Lower values were observed for India ($\hat{\rho}$ = 0.040) and Japan ($\hat{\rho}$ = 0.153), while higher reporting completeness was found for the United States ($\hat{\rho}$ = 0.416) and Italy ($\hat{\rho}$ = 0.492). These differences partly explain the variation in model errors and reflect known differences in surveillance intensity. Table 2 reports the complete comparison of forecasting performance across all five modeling approaches (SEIR baseline, Track A, Track B, the persistence baseline, and the pure ML model) for each country, covering MAE, RMSE, R^2, MAPE, high-incidence MAPE, and peak-timing error.

Table 2 shows that the hybrid SEIR-AI extensions did not consistently improve on the baseline model over the common held-out test window. Learning the transmission rate $\beta(t)$ with machine learning (Track A) produced RMSE values almost identical to SEIR in all four countries (5699.04 India, 32670.20 US, 19067.45 Italy, 90026.60 Japan). The learned transmission surface therefore converges to essentially the same degenerate flat pattern as the train-only SEIR fit once training-period information is exhausted. Residual learning (Track B) produced a small RMSE reduction relative to SEIR in all four countries (4957.53 vs. 5699.04 for India, 31252.23 vs. 32580.98 for the United States, 18532.09 vs. 19065.10 for Italy, and 89434.62 vs. 90191.22 for Japan). This modest and consistent gain suggests that residual correction can partially compensate for the mechanistic model's inability to track new waves, but the correction is far too small to bring Track B close to the persistence

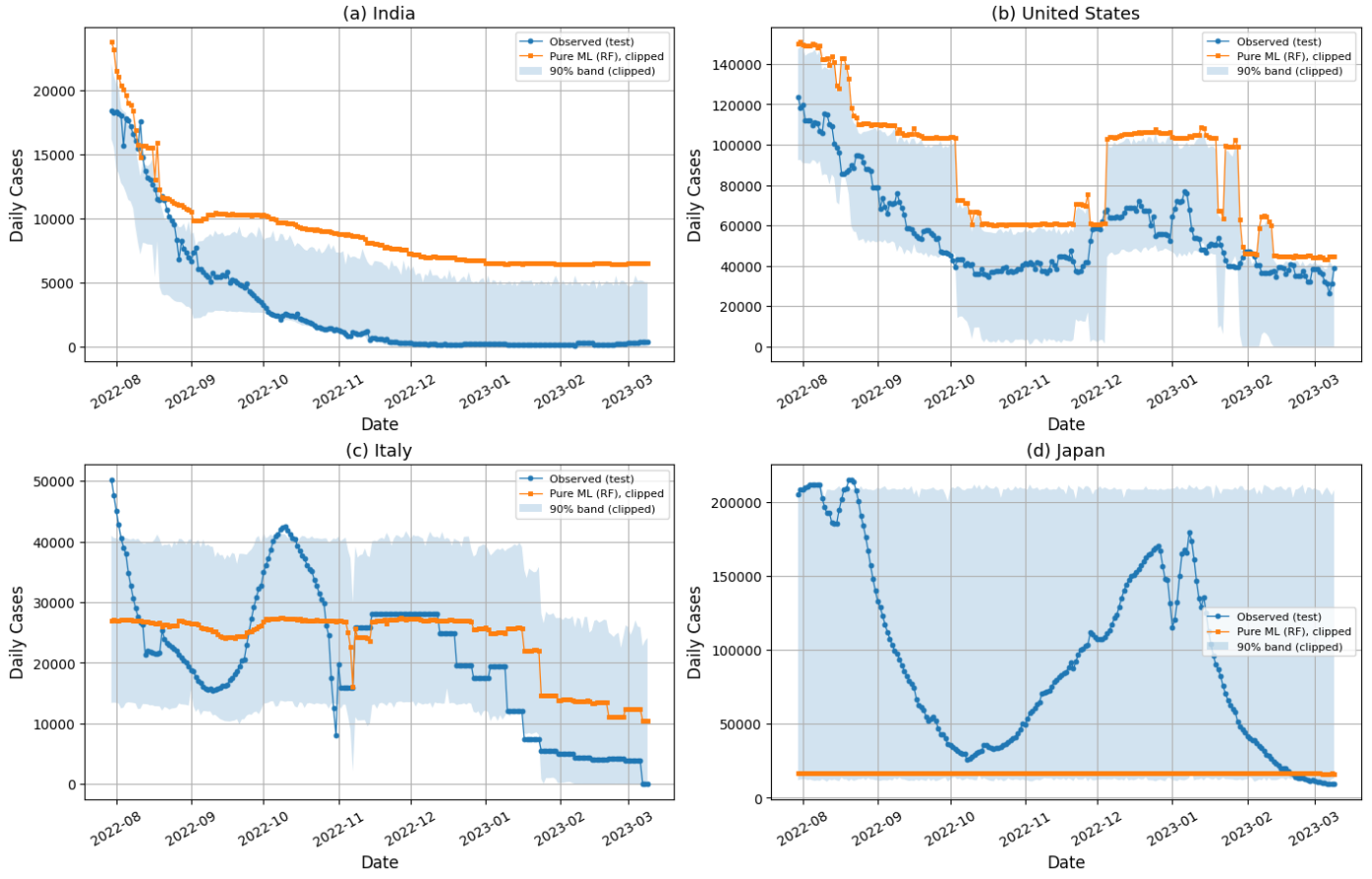

Figure 8: Observed versus predicted daily incidence during the held-out test period for the Random Forest forecasting model in India, the United States, Italy, and Japan. Shaded regions represent the 90% bootstrap-based prediction intervals, illustrating the uncertainty associated with the model forecasts.

baseline in any country.

Short-term forecasting experiments reinforced these findings. On the common 223-day held-out test window, a simple persistence baseline consistently outperformed the SEIR baseline, both hybrid extensions, and the Random Forest model across all four countries. To assess whether these differences were statistically meaningful, we applied the Diebold–Mariano test [32] to compare squared forecast errors between the persistence baseline and each other model, for all four countries (the train-only SEIR fit no longer produces the numerical instability present in an earlier, leakage-containing version of this analysis, so India is included in this comparison).

The persistence baseline produced significantly lower squared errors than the SEIR baseline in all four countries (India $DM = 6.36$; United States $DM = 13.51$; Italy $DM = 12.81$; Japan $DM = 14.63$; all $p < 0.001$), than Track A in all four countries (all $p < 0.001$), than Track B in all four countries (all $p < 0.001$), and than Random Forest in all four countries (all $p < 0.001$). This indicates that the autocorrelated structure of short-term COVID-19 incidence is captured substantially better by a naive one-step-ahead carry-forward forecast than by any of the mechanistic, hybrid, or machine-learning approaches examined in this study, once evaluated under a genuinely held-out, leakage-free test window.

For India, the persistence model achieved an RMSE of 430.10 and MAPE of 6.81%, whereas the Random Forest model produced substantially higher errors (RMSE = 6142.22, MAPE = 2010.63%). Similar degradation in machine-learning performance was observed for the United States, Italy, and Japan, with Random Forest peak-timing errors of 120 days for Italy and 184 days for Japan, indicating a limited ability to anticipate epidemic turning points; by contrast, the one-step-ahead persistence baseline's peak-timing errors did not exceed 1 day in any country (Table 2).

To mitigate distortion from very small incidence values, MAPE was evaluated during high-incidence periods, restricted to days with observed cases ≥ 100. Under this clinically relevant criterion, the persistence baseline achieved the lowest high-incidence MAPE for Japan (4.01%), compared to 89.68% for SEIR, 89.38% for Track A, 89.33% for Track B, and 71.39% for the pure machine-learning forecast on the same country; the complete per-country high-incidence MAPE values for all four study countries are reported in Table 2.

Taken together, the results demonstrate that AI-based predictive modeling is most effective when used to complement, rather than replace, mechanistic epidemic frameworks. The

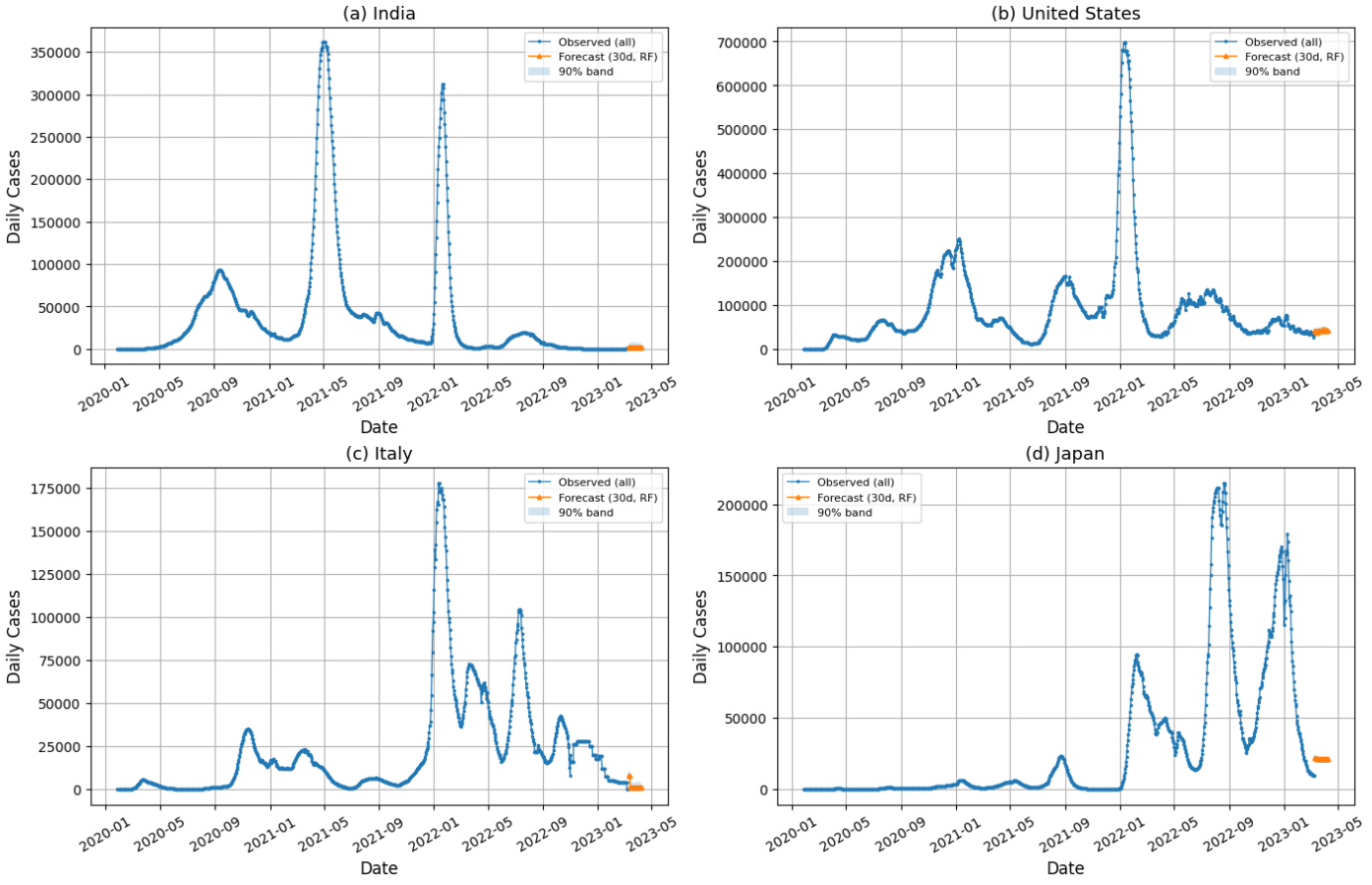

Figure 9: Thirty-day ahead forecasts of daily COVID-19 incidence generated by the Random Forest forecasting model for India, the United States, Italy, and Japan. Shaded regions represent the 90% bootstrap-based prediction intervals, illustrating the uncertainty associated with the long-term forecasts.

SEIR model provided the strongest baseline for stability and interpretability, and the simple persistence baseline outperformed the pure Random Forest model across all four countries on the held-out test window, indicating that the autocorrelated structure of short-term incidence was better captured by a naive carry-forward forecast than by the Random Forest model's broader feature set. The hybrid AI components (Track A and Track B) did not consistently improve short-term tracking accuracy and, in several cases, introduced additional instability. These findings refine the role of AI in infectious disease prediction by identifying its practical strengths, limitations, and appropriate scope of application.

5. Conclusion

A hybrid epidemic forecasting framework was developed that combines a mechanistic SEIR model with machine-learning components. The baseline SEIR model successfully reconstructed the main COVID-19 waves within its training window and provided an interpretable representation of transmission dynamics; however, once evaluated honestly on a genuinely held-out test window with $\beta(t)$ frozen at its last training-fitted value, its forecasting performance – and that of both hybrid extensions – was substantially weaker than a simple one-step-ahead persistence baseline in every country studied. Machine-learning extensions were explored as a means of introducing data-driven flexibility into the mechanistic framework. However, these extensions did not consistently improve forecasting accuracy and, for several countries, increased prediction error relative to the SEIR baseline; their performance depended heavily on data quality and the stability of reporting patterns.

The results indicate that the role of artificial intelligence in epidemic forecasting is not to replace mechanistic models. Instead, AI can complement these models by providing adaptive and diagnostic capabilities. The hybrid SEIR-AI framework helps identify when data-driven learning improves short-term responsiveness and when it may introduce instability under noisy surveillance conditions. This perspective connects AI-based predictive modeling with practical public health forecasting needs and provides a useful foundation for future hybrid epidemic forecasting research. The present study deliberately restricts its machine-learning features to lagged incidence and the effective reproduction number in order to isolate the comparative behaviour of mechanistic, machine-learning, and hybrid approaches under a controlled and interpretable feature set. Several natural extensions of this work are identified for

future research. First, incorporating additional external drivers of transmission – such as meteorological conditions, human mobility patterns, non-pharmaceutical intervention indices, and vaccination coverage – into the hybrid SEIR–machine-learning framework is a direct and important next step and is the focus of ongoing follow-up work by the authors. Second, replacing the current point-estimate parameter fitting procedure with a fully Bayesian inference framework (e.g., Markov Chain Monte Carlo or variational inference) would provide principled uncertainty quantification for $\beta(t)$, ρ, and ℓ, and would allow propagation of parameter uncertainty into forecast intervals in a statistically coherent manner. Third, replacing the Random Forest model with more expressive neural architectures – such as long short-term memory (LSTM) networks, temporal convolutional networks (TCN), or transformer-based sequence models – may better capture long-range temporal dependencies in epidemic incidence and could be incorporated as the machine-learning component within the Track A or Track B hybrid frameworks. Finally, extending the evaluation framework to a wider range of infectious diseases beyond COVID-19 (e.g., seasonal influenza, dengue, or mpox) and to more countries with heterogeneous surveillance systems would provide a more comprehensive assessment of the generalizability of the findings reported here.

Data availability

The COVID-19 case data used in this study were obtained from the Johns Hopkins University CSSE COVID-19 Data Repository, publicly available at https://github.com/CSSEGISandData/COVID-19, accessed on 5 July 2025. The specific file used was `time_series_covid19_confirmed_global.csv`. Processed datasets and analysis code generated during this study are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Funding

The authors received no specific funding from any public, commercial, or not-for-profit organisation for this research.

References

[1] W. O. Kermack & A. G. McKendrick, "A contribution to the mathematical theory of epidemics", Proceedings of the Royal Society of London. Series A **115** (1927) 700. https://doi.org/10.1098/rspa.1927.0118.

[2] R. Gupta, G. Pandey, P. Chaudhary & S. Pal, "SEIR and regression model based COVID-19 outbreak predictions in India", medRxiv (2020) 2020.04.01.20049825. https://doi.org/10.1101/2020.04.01.20049825.

[3] R. C. Poonia, A. K. J. Saudagar, A. Altameem, M. Alkhathami, M. B. Khan & M. H. A. Hasanat, "An enhanced SEIR model for prediction of COVID-19 with vaccination effect", Life **12** (2022) 647. https://doi.org/10.3390/life12050647.

[4] B. Ivorra, M. R. Ferrández, M. Vela-Pérez & A. M. Ramos, "Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections: the case of China", Communications in Nonlinear Science and Numerical Simulation **88** (2020) 105303. https://doi.org/10.1016/j.cnsns.2020.105303.

[5] Y. Mohamadou, A. Halidou & P. T. Kapen, "A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19", Applied Intelligence **50** (2020) 3913. https://doi.org/10.1007/s10489-020-01770-9.

[6] M. S. Rahman & A. H. Chowdhury, "A data-driven eXtreme gradient boosting machine learning model to predict COVID-19 transmission with meteorological drivers", PLOS ONE **17** (2022) e0273319. https://doi.org/10.1371/journal.pone.0273319.

[7] R. Chandra, A. Jain & D. S. Chauhan, "Deep learning via LSTM models for COVID-19 infection forecasting in India", PLOS ONE **17** (2022) e0262708. https://doi.org/10.1371/journal.pone.0262708.

[8] S. M. Shakeel, N. S. Kumar, P. P. Madalli, R. Srinivasaiah & D. R. Swamy, "COVID-19 prediction models: a systematic literature review", Osong Public Health and Research Perspectives **12** (2021) 215. https://doi.org/10.24171/j.phrp.2021.0100.

[9] R. Qasrawi, G. Issa, S. Thwib, R. AbuGhoush, M. Amro, R. Ayyad, S. Vicuña, E. Badran, Y. Khader, R. Al Qutob, F. Al Bakri, H. Trigui, E. Sokhn, E. Musa & J. D. Kong, "The role of machine learning in infectious disease early detection and prediction in the MENA region: a systematic review", Informatics in Medicine Unlocked **56** (2025) 101651. https://doi.org/10.1016/j.imu.2025.101651.

[10] M. Ala'raj, M. Majdalawieh & N. Nizamuddin, "Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections", Infectious Disease Modelling **6** (2021) 98. https://doi.org/10.1016/j.idm.2020.11.007.

[11] R. Vega, L. Flores & R. Greiner, "SIMLR: machine learning inside the SIR model for COVID-19 forecasting", Forecasting **4** (2022) 72. https://doi.org/10.3390/forecast4010005.

[12] J. Gao, R. Sharma, C. Qian, L. M. Glass, J. Spaeder, J. Romberg, J. Sun & C. Xiao, "STAN: spatio-temporal attention network for pandemic prediction using real-world evidence", Journal of the American Medical Informatics Association **28** (2021) 733. https://doi.org/10.1093/jamia/ocaa322.

[13] M. Z. Sayed-Ahmed, M. A. Hossain, S. Limkar, H. S. El-Bahkiry, N. Alam & S. T. Amin, "Mathematical modelling and deep learning techniques for predicting cardiovascular disease", Panamerican Mathematical Journal **34** (2024) 230. https://doi.org/10.52783/pmj.v34.i4.1880.

[14] B. Rai, A. Shukla & L. K. Dwivedi, "COVID-19 in India: predictions, reproduction number and public health preparedness", medRxiv (2020) 2020.04.09.20059261. https://doi.org/10.1101/2020.04.09.20059261.

[15] T. Phan, S. Brozak, B. Pell, A. Gitter, A. Xiao, K. D. Mena, Y. Kuang & F. Wu, "A simple SEIR-V model to estimate COVID-19 prevalence and predict SARS-CoV-2 transmission using wastewater-based surveillance data", Science of The Total Environment **857** (2023) 159326. https://doi.org/10.1016/j.scitotenv.2022.159326.

[16] D. Bolatova, S. Kadyrov & A. Kashkynbayev, "Mathematical modeling of infectious diseases and the impact of vaccination strategies", Mathematical Biosciences and Engineering **21** (2024) 7103. https://doi.org/10.3934/mbe.2024314.

[17] H. Verma, S. Mandal & A. Gupta, "Temporal deep learning architecture for prediction of COVID-19 cases in India", Expert Systems with Applications **195** (2022) 116611. https://doi.org/10.1016/j.eswa.2022.116611.

[18] J. Luo, Z. Zhang, Y. Fu & F. Rao, "Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms", Results in Physics **27** (2021) 104462. https://doi.org/10.1016/j.rinp.2021.104462.

[19] M. Saqib, "Forecasting COVID-19 outbreak progression using hybrid polynomial–Bayesian ridge regression model", Applied Intelligence **51** (2020) 2703. https://doi.org/10.1007/s10489-020-01942-7.

[20] B. Fatimah, P. Aggarwal, P. Singh & A. Gupta, "A comparative study for predictive monitoring of COVID-19 pandemic", Applied Soft Computing **122** (2022) 108806. https://doi.org/10.1016/j.asoc.2022.108806.

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

[22] G. O. Odekina, A. F. Adedotun & O. F. Imaga, "Modeling and forecasting the third wave of COVID-19 incidence rate in Nigeria using vector autoregressive model approach", Journal of the Nigerian Society of Physical Sciences **4** (2022) 117. https://doi.org/10.46481/jnsps.2022.431.

[23] D. O. Oyewola, E. G. Dada, J. N. Ndunagu, T. A. Umar & S. A. Akinwunmi, "COVID-19 risk factors, economic factors, and epidemiological factors nexus on economic impact: machine learning and structural equation modelling approaches", Journal of the Nigerian Society of Physical Sciences **3** (2021) 395. https://doi.org/10.46481/jnsps.2021.173.

[24] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim & U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images", Computers in Biology and Medicine **121** (2020) 103792. https://doi.org/10.1016/j.compbiomed.2020.103792.

[25] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang & W. Ji, "Sensitivity of chest CT for COVID-19: comparison to RT-PCR", Radiology **296** (2020) E115. https://doi.org/10.1148/radiol.2020200432.

[26] D. J. McDonald, J. Bien, A. Green, A. J. Hu, N. DeFries, S. Hyun, N. L. Oliveira, J. Sharpnack, J. Tang, R. Tibshirani, V. Ventura, L. Wasserman & R. J. Tibshirani, "Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction?", Proceedings of the National Academy of Sciences **118** (2021) e2111453118. https://doi.org/10.1073/pnas.2111453118.

[27] H. Alkhalefah, D. Preethi, N. Khare, M. H. Abidi & U. Umer, "Deep learning infused SIRVD model for COVID-19 prediction: XGBoost-SIRVD-LSTM approach", Frontiers in Medicine **11** (2024) 1427239. https://doi.org/10.3389/fmed.2024.1427239.

[28] M. U. G. Kraemer *et al.*, "Artificial intelligence for modelling infectious disease epidemics", Nature **638** (2025) 623. https://doi.org/10.1038/s41586-024-08564-w.

[29] R. Vaishya, M. Javaid, I. H. Khan & A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic", Diabetes & Metabolic Syndrome: Clinical Research & Reviews **14** (2020) 337. https://doi.org/10.1016/j.dsx.2020.04.012.

[30] E. Dong, H. Du & L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time", The Lancet Infectious Diseases **20** (2020) 533. https://doi.org/10.1016/S1473-3099(20)30120-1.

[31] World Bank, "Population, total" (indicator SP.POP.TOTL), World Development Indicators, accessed 5 July 2025. https://data.worldbank.org/indicator/SP.POP.TOTL.

[32] F. X. Diebold & R. S. Mariano, "Comparing predictive accuracy", Journal of Business & Economic Statistics **13** (1995) 253. https://doi.org/10.1080/07350015.1995.10524599.