Machine Learning and Probabilistic Approaches for Forecasting Infectious Disease Transmission and Cases

Md Sakhawat Hossain,[1,2*] Ravi Goyal,[3] Natasha K Martin,[3] Victor DeGruttola,[4] Tanvir Ahammed,[1,2] Christopher McMahan,[2,5†] Lior Rennert,[1,2*†]

1 Department of Public Health Sciences, Clemson University, Clemson, SC, USA

2 Center for Public Health Modeling and Response, Clemson University, Clemson, SC, USA

3 Division of Infectious Diseases & Global Public Health, University of California San Diego, La Jolla, CA, USA

4 Division of Biostatistics, Herbert Wertheim School of Public Health and Longevity Science, University of California San Diego, San Diego, California, USA

5 School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, USA

† Senior Authors

* liorr@clemson.edu, mdsakhh@clemson.edu

*** Correspondence**
Dr. Lior Rennert, Center for Public Health Modeling and Response, Department of Public Health Sciences, Clemson University, Clemson, SC 29634, USA
Email: liorr@clemson.edu

Dr. Md Sakhawat Hossain, Center for Public Health Modeling and Response, Department of Public Health Sciences, Clemson University, Clemson, SC 29634, USA
Email: mdsakhh@clemson.edu

Abstract

Objectives

Forecasting the effective reproductive number (R_t) and infection case counts is critical for guiding public health responses. We developed a machine learning and probabilistic forecasting framework to predict R_t and daily COVID-19 cases, respectively, across South Carolina counties, with the flexibility to generalize to other infectious diseases.

Methods

We first estimated R_t using the EpiNow2 R package, which incorporates Bayesian time-series modeling and accounts for reporting delay and incubation period. These initial estimates were refined using spatial covariate-adjusted smoothing through the Integrated Nested Laplace Approximation (INLA). We then generated R_t forecasts using an ensemble of linear regression, random forest, and XGBoost models. Daily case forecasts were obtained by linking R_t trajectories with historical case data via a Poisson model.

Results

This ensemble-based approach outperformed EpiNow2 across different forecast horizons (7-day, 14-day, and 21-day). In the first forecast period (November 11, 2020 – February 02, 2021), the ensemble achieved a median PA of 96.5% (IQR: 95.4% – 97.1%) for 7-day horizon R_t forecast, compared to 87.0% (IQR: 84.4% – 89.4%) from EpiNow2. In the second period (December 11, 2022 – March 04, 2023), the ensemble attained a 93.0% median PA for R_t forecast (IQR: 90.8% – 95.4%), while EpiNow2 reached 86.8% (IQR: 82.5% – 89.2%). Similar trends were observed for case forecasts, with the ensemble model demonstrating improved performance.

Conclusion

This study presents a flexible forecasting framework that integrates Bayesian estimation, spatial smoothing, and ensemble machine learning to improve the accuracy of COVID-19 transmission and case forecasts. The approach enhances epidemic forecasting performance and offers scalable tools to support data-driven public health preparedness and response.

Keywords: Effective reproductive number, Infectious disease modeling, Machine learning, Forecasting, COVID-19.

1 Introduction

The COVID-19 pandemic has significantly reshaped global priorities in public health, focusing on the importance of data-driven models to predict and manage disease spread. A crucial parameter in understanding and forecasting infectious diseases is the effective reproductive number (R_t), which indicates the average number of secondary cases caused by an infected individual at a given time t. This parameter is widely utilized to monitor and assess outbreak progression and has played a critical role in informing public health interventions and policy decisions during the COVID-19 pandemic [1,2]. Values greater than 1 indicate disease transmission is increasing, while values below 1 indicate the disease is contracting [3,4].

Accurately estimating and forecasting R_t is crucial for assessing the transmissibility of infectious disease and guiding public health interventions, including social distancing, mask mandates, vaccination strategies, and resource allocation [1,5–8]. The R_t estimates can be linked to changes in policy, population behavior and immunity, pathogen evolution, and other factors [1,8–10].

Various methods have been developed to estimate R_t, including those capable of real-time monitoring, each with its unique strengths and limitations [1,5,11–18]. EpiNow2 enhanced estimation accuracy by incorporating generation time intervals, incubation period, and reporting delays [1,11]. Most methods focus on retrospective estimation of R_t rather than future prediction[1,5,12,14,16,18–21]. While epidemic forecasting is a growing field, much of it focused on predicting case counts and other metrics, with relatively fewer studies explicitly forecasting the future course of R_t[22–26].

Machine learning has significantly advanced infectious disease modeling, aiding in outbreak detection, diagnosis, severity prediction, and forecasting of R_t and case counts [27–30].

Supervised learning techniques, such as random forest and gradient boosting, have been widely used for predicting disease incidence and hospitalizations, often outperforming traditional regression methods [31,32]. Combining forecasts from multiple models has emerged as a powerful approach for epidemic forecasting. Ensemble methods that combine outputs from multiple models have demonstrated improved robustness [33], particularly in scenarios with inconsistent or incomplete data [30].

In this study, we develop a framework for forecasting the effective reproductive number (R_t) and, subsequently, daily case counts. First, we estimate R_t using the *R software package* EpiNow2. Next, we apply a spatial (covariate-adjusted) smoothing through INLA method to refine the R_t estimates [34]. We then employ regression and machine learning models to forecast R_t using both the initial and smooth estimates. Finally, a stochastic Poisson process is implemented to generate COVID-19 case counts based on R_t forecasts, utilizing historical data and the disease serial interval distribution.

2 Materials and Methods

We develop a forecasting framework to predict the effective reproductive number (R_t) and daily COVID-19 case counts. We apply our method to predict these quantities for 46 South Carolina (SC) counties. R_t is initially estimated using the EpiNow2 R package, which applies Bayesian time-series modeling, accounting for reporting delay and incubation period. These estimates are then refined using a spatial (covariate-adjusted) smoothing process through the Integrated Nested Laplace Approximation (INLA) method.

To forecast R_t, we employ linear regression (Reg), XGBoost (XGB), and Random Forest (RF) models, leveraging lagged values of the R_t estimates and temporal features such as the day of the

year. A recursive approach is applied to generate multi-step forecasts. An ensemble-based forecasting approach is then developed by combining outputs from individual models.

For COVID-19 case forecasting, we use the forecast together with historical case data and the serial interval distribution. Forecasts are generated recursively, incorporating updated case counts at each step to refine predictions. **Figure 1** presents the workflow diagram of and daily COVID-19 case forecasts.

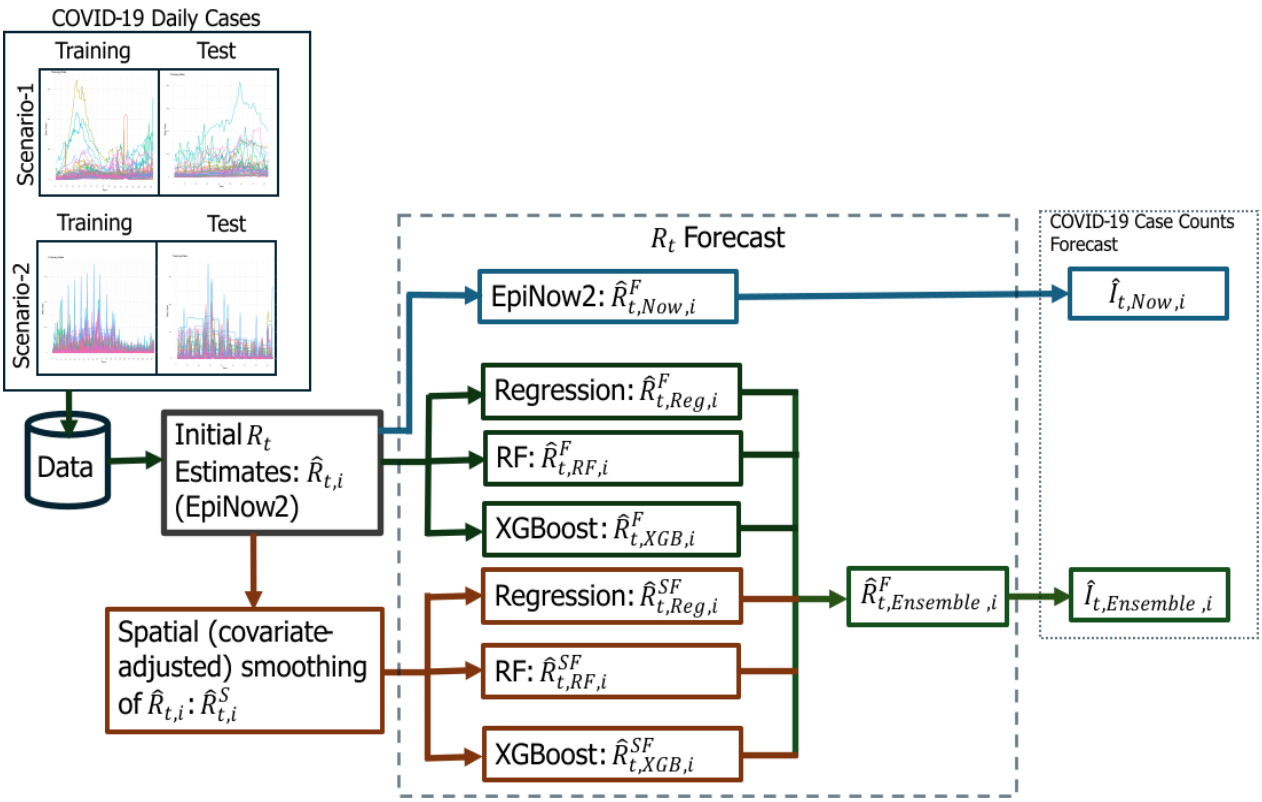

Figure 1. Flow diagram of the forecasting framework for effective reproductive number and COVID-19 case counts at the county level in SC. Scenario-1 includes the data from June 01, 2020, to November 10, 2020 (Training) and November 11, 2020, to February 02, 2021 (Test, forecast period), while Scenario-2 covers data from May 01, 2022, to December 10, 2022 (Training) and December 11, 2022, to March 04, 2023 (Test, forecast period). Initially, is estimated using the EpiNow2 package, followed by spatial (covariate-adjusted) smoothing via the INLA model. The initial and smoothed estimates are denoted as and , respectively. forecasting is performed

using Reg, RF, and XGB on both initial and spatially smoothed estimates, followed by an ensemble-based approach for final predictions. The ensemble forecasted R_t is denoted as $\hat{R}^F_{t,Ensemble,i}$, while $\hat{I}_{t,Ensemble,i}$ and $\hat{I}_{t,Now,i}$ represent COVID-19 case count forecasts from the ensemble approach and EpiNow2, respectively.

We perform the initial estimation of R_t for the period June 01, 2020, to March 04, 2023. The forecasts for both R_t and COVID-19 daily case counts are generated for the time periods November 11, 2020 – February 02, 2021 (Scenario-1) and December 11, 2022 – March 04, 2023 (Scenario-2). The forecasting approach involves 7-day, 14-day, and 21-day horizon predictions over 84 days period using a rolling window method. All models are trained on R_t estimates from June 01, 2020, to November 10, 2020, for forecasting under Scenario-1, and from May 01, 2022, to December 10, 2022, for Scenario-2. Then we generate R_t forecasts through an ensemble-based approach that combines predictions from all individual models. To benchmark the performance of our models, we also produce R_t and case forecasts using the EpiNow2 R package, however, these forecasts are used only for comparison purposes and are not included in the ensemble.

2.1 Data sources

We utilize publicly available county-level COVID-19 daily case data for SC from the New York Times (NYT) GitHub repository [35]. The demographic variables, including population by age, sex, race/ethnicity, employment, and insurance coverage are used. These data are sourced from the United States Census Bureau website [36]. The social vulnerability index (SVI) is obtained from the Agency for Toxic Substances and Disease Registry website [37]. Because daily case reports show irregular updates and retrospective adjustments, we use a 7-day moving average to smooth short-term fluctuations in the data.

2.2 Initial estimation and spatial (covariate-adjusted) smoothing of R_t

We first utilize the EpiNow2 R package to estimate the effective reproductive number, R_t. It incorporates cases by reported date, generation time interval, delay distributions such as incubation and reporting delays, more details can be found at Abbott et al. [11]. The incubation period distribution and the reporting delay distribution are modeled using gamma distribution. For our analysis, we consider a mean incubation period of 5 days (95% CI, 4.94-5.06) with a standard deviation of 2.4 days [38,39], a mean serial interval distribution of 4.7 days with a standard deviation of 2.9 days [40], and a mean reporting delay of 3.2 days [41]. The initial estimate of R_t using EpiNow2 at a given time t and for a specific region i is denoted by $\hat{R}_{t,i}$. To spatially smooth R_t estimates, a spatial (covariate-adjusted) Bayesian model is employed, as described in [42]. The model incorporates sociodemographic covariates, including age, race, employment, insurance status, SVI, and normalized median income. We denote the smoothed estimates by $\hat{R}^S_{t,i}$.

2.3 Forecasting effective reproductive number (R_t)

In this section, we outline the methodology for forecasting the effective reproductive number utilizing both estimates $\hat{R}_{t,i}$ and $\hat{R}^S_{t,i}$ across M regions. These forecasts are denoted as $\hat{R}^F_{t,i}$ and $\hat{R}^{SF}_{t,i}$ for region i at time t. For generalization, we represent these time series as $y_{t,i}$, corresponding to either $\hat{R}_{t,i}$ or $\hat{R}^S_{t,i}$. We consider the lagged estimates as predictors and incorporate temporal features, such as the day of the year(D_t).

The time series data for region i is $y_{1,i}, y_{2,i}, \dots, y_{T,i}$, where T is the most recent date used for forecasting. The goal is to forecast $y_{T+h,i}$ for $h \in \{1, 2, \dots, H\}$, where H is the forecasting

horizon. For each region i, the features are constructed using p lagged values of the time series and the day of the year corresponding to the forecasted time point (D_{t+h}).

The feature matrix for all regions is constructed as below:

$$\boldsymbol{X} = \begin{pmatrix} y_{1,1} & y_{2,1} & \cdots & y_{p,1} & D_{p+1,1} \\ y_{2,1} & y_{3,1} & \cdots & y_{p+1,1} & D_{p+2,1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{T-p,1} & y_{T-p+1,1} & \cdots & y_{T-1,1} & D_{T,1} \\ y_{1,2} & y_{2,2} & \cdots & y_{p,2} & D_{p+1,2} \\ y_{2,2} & y_{3,2} & \cdots & y_{p+1,2} & D_{p+2,2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{T-p,M} & y_{T-p+1,M} & \cdots & y_{T-1,M} & D_{T,M} \end{pmatrix}$$

The response vector is denoted as:

$$\boldsymbol{y} = \begin{pmatrix} y_{p+1,1} & y_{p+2}, 1 & \cdots & y_{T,1} & y_{p+1,2} & y_{p+2,2} & \cdots & y_{T,M} \end{pmatrix}'.$$

Then the model is expressed as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is response vector of $(T-p) \times M$ elements, $\mathbf{X}$ is features matrix of $(T-p) \times M$ rows and $(p+1)$ columns, $\boldsymbol{\beta}$ is the column vector of model parameters, and ϵ is the residual error term.

To determine the optimal number of lagged features (p^*) for the final model, we train the model using values for $p \in \{2, 3, \ldots, P\}$. RF and XGB models are implemented in R using the 'caret', 'randomForest', and 'xgboost' packages. For the XGB model, we focus on tuning the learning rate (η), L_2 regularization penalty (λ), and maximum tree depth, while for RF model, the number of trees (ntree) and the maximum number of terminal nodes (maxnodes) are tuned. All the optimizations are based on achieving the maximum PA values. After identifying the optimal

lag parameter (p^*) and hyperparameters, the final model is trained using the data up to $y_{T,i}$ for each region i. Once the final model is fitted, h-step horizon forecasts are generated through a recursive forecasting method.

The steps are as follows:

1. Feature construction: At time $T+1$, form the feature vector using the last p^*observations and day of the year:

$$\boldsymbol{X}_{T+1,i} = [y_{T-p^*+1,i} \quad y_{T-p^*+2,i} \quad \cdots \quad y_{T-1,i} \quad y_{T,i} \quad D_{T+1,i}].$$

2. Prediction: Obtain the forecast $\hat{y}_{T+1,i}$ at time $T+1$ and location i from the fitted Reg, RF, and XGB models.
3. Update: Append $\hat{y}_{T+1,i}$ to the feature vector, discard the oldest lag, and construct

$$\boldsymbol{X}_{T+2,i} = [y_{T-p^*+2,i} \quad y_{T-p^*+3,i} \quad \cdots \quad y_{T,i} \quad \hat{y}_{T+1,i} \quad D_{T+2,i}].$$

4. Iteration: Repeat steps 2-3 to generate forecasts $\hat{y}_{T+1,i}$, $\hat{y}_{T+2,i}$, …, $\hat{y}_{T+h,i}$ over horizon h.

Forecasts based on the initial estimates $\hat{R}_{t,i}$ are denoted as $\hat{R}^F_{t,Reg,i}$, $\hat{R}^F_{t,RF,i}$, and $\hat{R}^F_{t,XGB,i}$, while those based on the spatially smoothed estimates $\hat{R}^S_{t,i}$ are denoted as $\hat{R}^{SF}_{t,Reg,i}$, $\hat{R}^{SF}_{t,RF,i}$, and $\hat{R}^{SF}_{t,XGB,i}$. To combine these results, we implement an ensemble forecasting framework in which each model's weight is determined from its relative percentage agreement (PA) value, with PA computed using equation (1) in Section 2.5. The ensemble forecast is denoted as $\hat{R}^F_{t,Ensemble,i}$.

2.4 Forecasting COVID-19 daily cases

To forecast the COVID-19 daily cases, we utilize a probabilistic framework that incorporates the serial interval distribution, historical case data, and forecasted effective reproductive number, $\hat{R}^F_{t,i}$. We use $\hat{R}^F_{t,i}$ as a general term to describe the case forecasting steps; however, case

forecasts are generated for each individual R_t forecast. The corresponding case are presented as $\hat{I}_{t,Reg,i}$, $\hat{I}_{t,RF,i}$, $\hat{I}_{t,XGB,i}$, $\hat{I}^S_{t,Reg,i}$, $\hat{I}^S_{t,RF,i}$, $\hat{I}^S_{t,XGB,i}$ and $\hat{I}_{t,Ensemble,i}$.

The serial interval, representing the time between symptom onset in a primary case and secondary case, is assumed to follow a gamma distribution. This distribution is used to compute the total infectiousness ($\Lambda_{t,i}$) for each day t at region i, which quantifies the potential to generate new infections from previously observed cases in s days. Total infectiousness is calculated as:

$$\Lambda_{t,i} = \sum_{k=1}^{s} I_{t-k,i} \cdot p_k$$

where $I_{t-k,i}$ is the number of observed cases on day $t-k$ at location i, p_k is the probability that a primary case generates a secondary case between $k-1$ and k days. The probabilities p_k represent relative probability mass function of the generation time distribution, which is typically modeled using a gamma or log-normal distribution[1,5].

Then the expected number of cases ($\mathrm{E}[I_{t,i}]$) is calculated as:

$$\mathrm{E}\left[I_{t,i}\right] = \Lambda_{t,i} \times \hat{R}^F_{t,i}$$

Daily case counts are then modeled as a Poisson random variable:

$$I_{t,i} \sim Poisson(\Lambda_{t,i} \times \hat{R}^F_{t,i})$$

For case forecasting, we simulate full h-step trajectories sequentially by appending the forecasted values at each time point to the historical data before generating the next step. At each time t, new cases are sampled from a Poisson distribution with mean $\Lambda_{t,i} \times \hat{R}^F_{t,i}$, where $\Lambda_{t,i}$ is computed from the updated series. This process is repeated N times to produce N forecast

trajectories, and the final case forecast at each horizon is obtained by averaging across the simulated trajectories.

2.5 Performance evaluation of R_t and case forecasts

Let $O_{t,i}$ denotes the observed values at region i and time t (e.g. daily cases, we do not know observed values for $R_{t,i}$, use estimates $\hat{R}^S_{t,i}$), $P_{t,i}$ represents the forecasted value at region i and time t. Then the evaluation metric is defined as follows:

Percentage Agreement (PA):

$$\mathrm{PA} = \frac{1}{T}\sum_{t=1}^{T}\frac{\min\,(O_{t,i},\ P_{t,i})}{\max\,(O_{t,i},\ P_{t,i})} \times 100 \tag{1}$$

For R_t forecast evaluation, we compute pointwise PA, meaning that the agreement is assessed for each time point t, followed by averaging the PA values over the entire forecasting period. For case forecast evaluation, we first aggregate both observed and forecasted daily case counts into consecutive 7-day totals over the forecast period, calculate weekly PA by comparing aggregated forecasts with observations, and average these weekly PA values across the entire forecast period.

We use EpiNow2 R package version 1.7.1, with seed = 100 for all analyses to ensure reproducibility. We perform 5-fold cross-validation for optimal lag selection and hyperparameter tuning. A detailed summary of the selected lagged features and model-specific hyperparameters is provided in the supplementary materials (**Table S1)**.

3 Results

3.1 Evaluation of R_t forecasting performance

The median percentage agreement (PA) across counties for ensemble-based forecasts was higher than that of EpiNow2 (**Table 1**). For Scenario-1, the ensemble method achieved a median PA of 96.5% (IQR: 95.4% – 97.1%) across counties for 7-day horizon forecasts, while EpiNow2 reached 87.0% (IQR: 84.4% – 89.4%). Similarly, in Scenario-2, the ensemble approach maintained a median PA of 93.0% (IQR: 90.8% – 95.4%) for 7-day horizon forecasts compared to a median PA of 86.8% (IQR: 82.5% – 89.2%) for EpiNow2.

Table 01. Forecast accuracy metrics for the effective reproductive number (R_t) in SC counties during two forecast periods: November 11, 2020 – February 02, 2021 (Scenario-1) and December 11, 2022 – March 04, 2023 (Scenario-2). Forecast models were trained with data from June 01, 2020, to November 10, 2020, for Scenario-1, and from May 01, 2022, to December 10, 2022, for Scenario-2. The forecasting approach employed a rolling window method for 7-day, 14-day, and 21-day horizon predictions over 84 days period. The table presents the median PA across counties along with the IQR of county PAs.

Forecast Method	Effective Reproductive Number Forecast: Percentage Agreement (PA), Median (IQR)					
	Forecast Period: November 11, 2020 – February 02, 2021 (Scenario-1)			Forecast Period: December 11, 2022 – March 04, 2023 (Scenario-2)		
	7-day horizon	14-day horizon	21-day horizon	7-day horizon	14-day horizon	21-day horizon
EpiNow2	87.0% (84.4% – 89.4%)	86.2% (81.5% – 89.5%)	87.6% (85.1% – 90.6%)	86.8% (82.5% – 89.2%)	86.7% (81.4% – 89.9%)	86.5% (79.5% – 89.6%)
Ensemble	96.5% (95.4% –	96.0% (94.8% –	94.1% (93.2% –	93.0% (90.8% –	92.6% (90.4% –	92.7% (90.8% –

	97.1%)	96.7%)	95.1%)	95.4%)	94.4%)	93.6%)

Table 1 provides a detailed comparison of forecasting accuracy for 7-day, 14-day, and 21-day horizons of forecasts for 84 days period during both Scenario-1 and Scenario-2. The ensemble-based approach outperformed EpiNow2 across all forecast horizons. The relative difference in the 7-day horizon forecast PA was 10.9% higher for the ensemble approach in Scenario-1 (96.5% vs 87.0%) and 7.1% higher in Scenario-2 (93.0% vs. 86.8%). The relative difference in 21-day horizon PA was 7.4% higher for the ensemble approach in Scenario-1 (94.1% vs 87.6%) and 7.2% higher in Scenario-2 (92.7% vs 86.5%).

In counties with higher COVID-19 case counts (**Figure 2,** left panel), both methods captured trends well, but EpiNow2 (red lines) exhibited more fluctuations, particularly in Horry and Lexington. In counties with lower COVID-19 case counts (**Figure 2,** right panel), the ensemble approach still performed better.

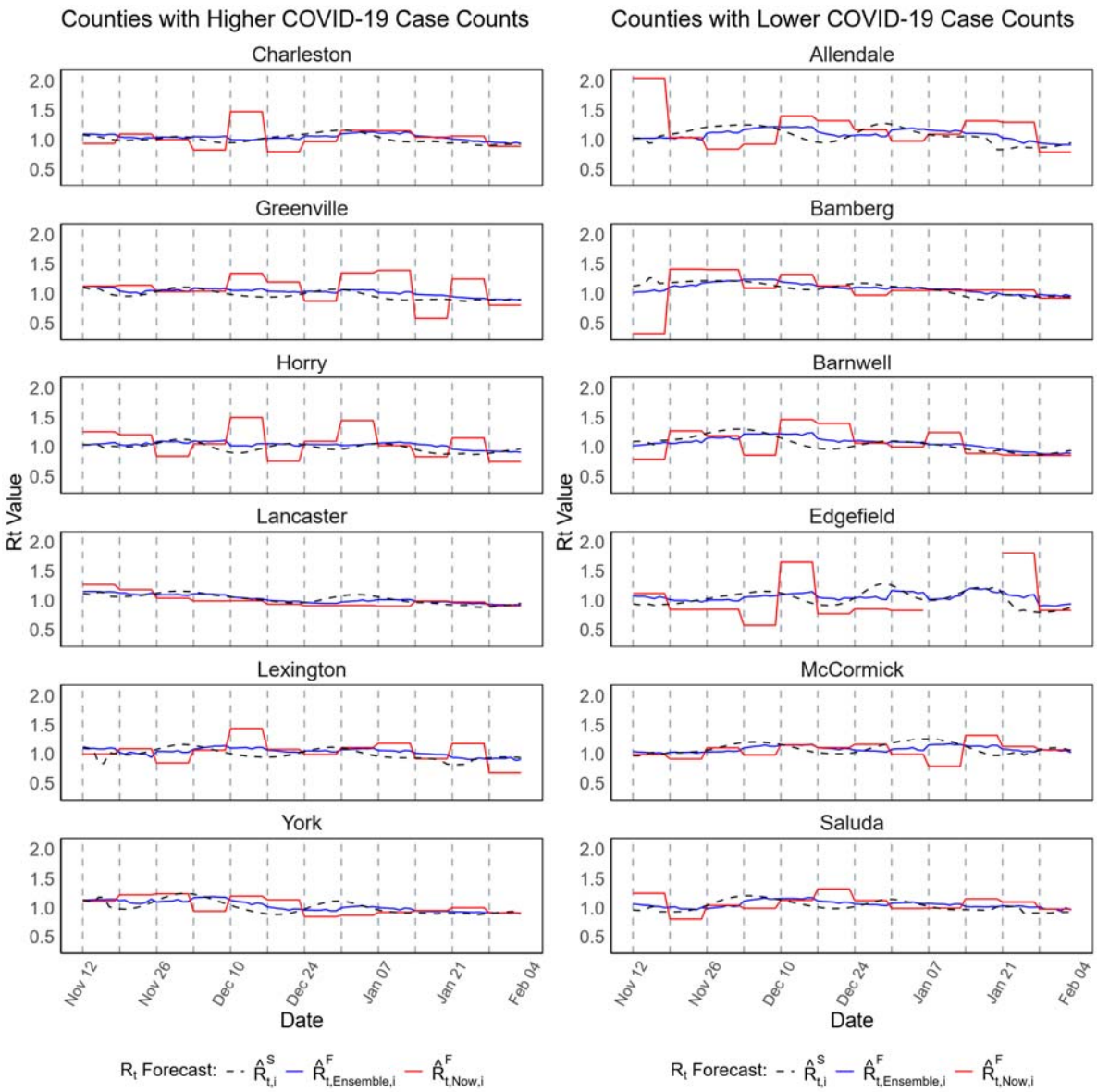

Figure 2. Forecast of at the county level in SC during Scenario-1 (November 11, 2020 – February 02, 2021). The left panel presents forecasts for six counties with higher COVID-19 case counts, while the right panel displays forecasts for six counties with lower case counts. The plots compare the ensemble-based forecasts (blue lines) with those generated using the **EpiNow2** R package (red lines), alongside the spatially (covariate-adjusted) smoothed estimates, (black dashed lines), where presents the county, denotes the time point (day), and "Now" refers to the **EpiNow2** method. Forecasts were generated using a rolling window approach for 7-day horizon predictions over 84 days period.

Across all counties, the ensemble approach provided higher agreement with observed trends, while EpiNow2 showed more variability, particularly in some counties with significant fluctuations (e.g., Horry and Edgefield). The supplementary materials provide further insights into forecast performance across all counties. **Figure S1** presents 7-day horizon R_t forecasts for all 46 counties, while **Figures S2** and **S3** display the 14-day and 21-day horizon forecasts, respectively, for Scenario-1. Similarly, **Figures S4-S6** provide the 7-day, 14-day, and 21-day forecasts for Scenario-2, reaffirming that the ensemble model outperformed EpiNow2.

Figure S7 presents the 7-day horizon forecasts of R_t for all 46 counties in SC under Scenario-1, where we compared the spatially smoothed forecast from EpiNow2 with the ensemble-based forecasts. Alongside the spatially smooth R_t estimates. **Figure S8** compares the forcasted R_t values with the initial R_t estimates, allowing assessment of model performance without spatial smoothing. **Table S2** summarizes the corresponding forecast accuracies for both Scenario-1 and Scenario-2. Across all horizons and both scenarios, the ensemble forecasts outperformed EpiNow2. For example, the ensemble achieved a median PA of 96.2% for 7-day horizon forecasts in Scenario-1 and 92.2% in Scenario-2. These correspond to approximately 12.3% and 4.1% higher relative performance compared to EpiNow2 forecasts, respectively.

3.2 Evaluation of COVID-19 case forecasting performance

The results indicated that the ensemble approach outperformed EpiNow2, particularly for short-term forecasts (7-day and 14-day horizons). In Scenario-1, the ensemble model achieved a median PA of 87.3% (IQR: 84.0% – 88.8%) for 7-day forecasts, slightly outperforming EpiNow2, which had a median PA of 83.8% (IQR: 81.6% – 86.9%). For longer horizons (21-day forecasts), the ensemble approach maintained a median PA of 82.6% (IQR: 80.0% – 86.5%),

while EpiNow2 showed a lower median PA of 67.6% (IQR: 61.3% – 75.1%), demonstrating that the ensemble method provided greater stability in extended forecasting periods.

A similar trend was observed in Scenario-2, where ensemble-based forecasts continued to outperform EpiNow2 across all forecast horizons. The relative improvement in PA was more evident for longer-term predictions, with the ensemble model achieving a median PA of 74.6% (IQR: 66.7% – 78.7%) for 21-day forecasts, compared to 66.2% (IQR: 57.4% – 73.2%) for EpiNow2. **Table 2** presents a detailed comparison of forecast accuracy metrics for 7-day, 14-day, and 21-day horizon predictions across both scenarios.

Table 02. Forecast accuracy metrics for COVID-19 daily case counts at the county level in SC during Scenario-1 (November 11, 2020 – February 2, 2021) and Scenario-2 (December 11, 2022 – March 4, 2023). Forecast accuracy was evaluated using PA with median and IQR across all 46 counties. Case forecasts were generated using the EpiNow2 R package and a probabilistic Poisson model based on the ensemble-based R_t forecasts ($\hat{R}^{F}_{t,Ensemble,i}$). Forecasts were conducted for 7-day, 14-day, and 21-day horizon predictions over 84 days period using a rolling window approach.

Forecast Method	COVID-19 Daily Cases Forecast: Percentage Agreement (PA), Median (IQR)					
	Forecast Period: November 11, 2020 – February 02, 2021 (Scenario-1)			Forecast Period: December 11, 2022 – March 04, 2023 (Scenario-2)		
	7-day horizon	14-day horizon	21-day horizon	7-day horizon	14-day horizon	21-day horizon
EpiNow2	83.8% (81.6% – 86.9%)	75.4% (72.4% – 78.7%)	67.6% (61.3% – 75.1%)	75.1% (67.5% – 79.1%)	73.5% (65.3% – 76.7%)	66.2% (57.4% – 73.2%)

Ensemble	87.3% (84.0% – 88.8%)	84.3% (82.6% – 87.6%)	82.6% (80.0% – 86.5%)	77.6% (72.3% – 80.4%)	77.2% (72.1% – 80.1%)	74.6% (66.7% – 78.7%)

The results indicate that the ensemble approach produced more stable and accurate case forecasts than EpiNow2, with differences varying by county and case count levels (**Figure 3**). In higher case count counties (Charleston, Greenville, Horry, Lancaster, Lexington, York), the ensemble model captured trends more smoothly, whereas EpiNow2 exhibited abrupt spikes and larger deviations from observed case counts, particularly in Horry and Lexington. In lower case count counties (Allendale, Bamberg, Barnwell, Edgefield, McCormick, Saluda), the performance gap between the two methods was more evident in certain areas, such as Edgefield.

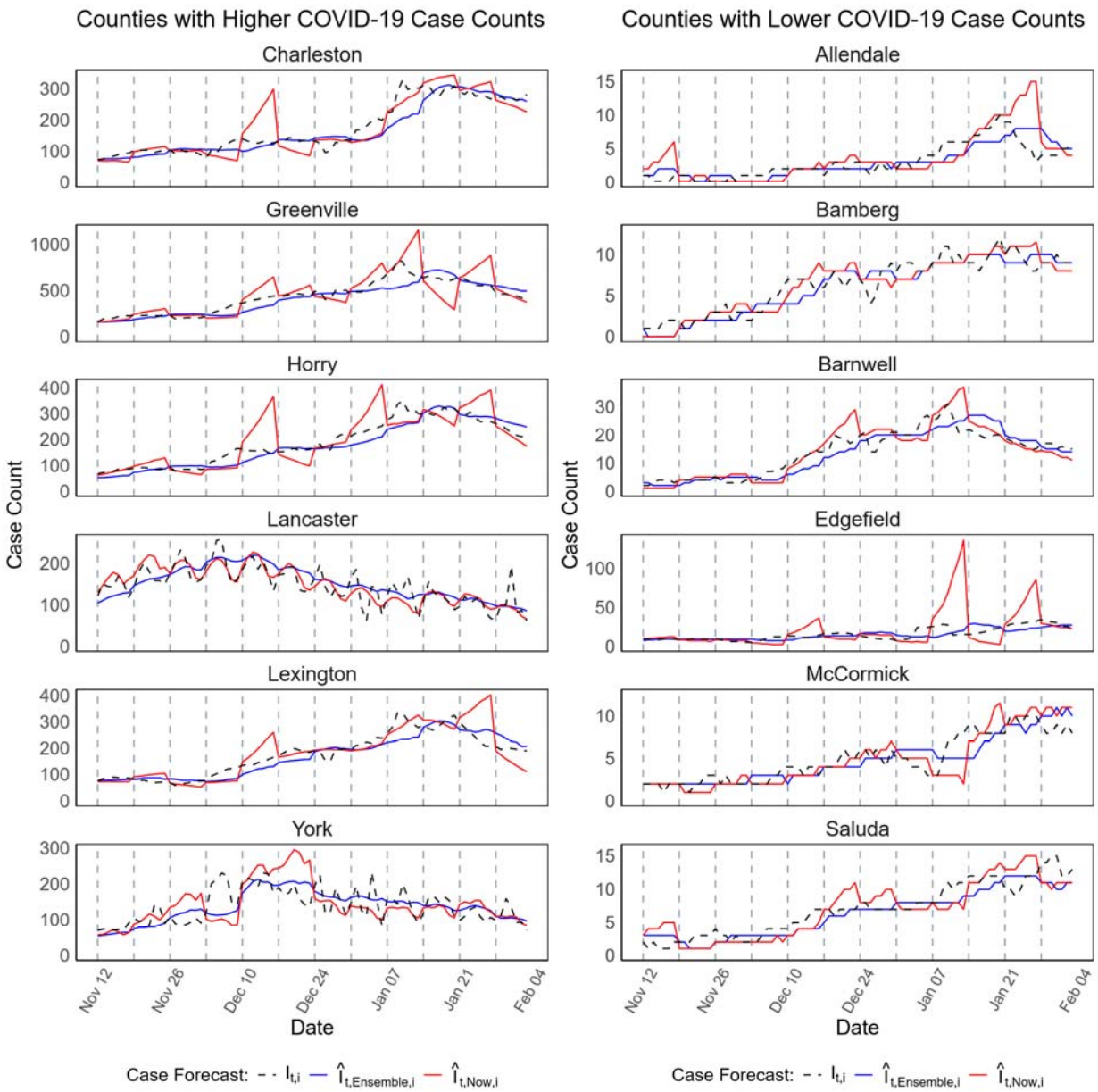

Figure 03. Forecast of COVID-19 case counts at the county level in SC during Scenario-1 (November 11, 2020 – February 02, 2021). The left panel presents forecasts for six counties with higher COVID-19 case counts, while the right panel displays forecasts for six counties with lower case counts. The plots compare EpiNow2 forecasts (red lines) and ensemble-based forecasts (blue lines) against the observed daily case counts (black dashed lines). The forecasts were generated for 7-day horizon predictions over 84 days period using a rolling window approach.

Overall, the results confirm that the ensemble approach provided more accurate and stable case forecasts across all time horizons and geographical regions. The supplementary materials provide a comprehensive summary of forecast performance across all counties. **Figures S9-S11** (Scenario-1) and **Figures S12-S14** (Scenario-2) present case forecasts for all counties at 7-day, 14-day, and 21-day horizon, respectively.

3.2 R_t and COVID-19 case forecasting performance of individual models

Table S3 indicates that RF models achieved higher median PA values compared to Reg and XGB models, especially when trained on spatially smoothed estimates. For instance, during Scenario-1, XGB model with spatial smoothing attained a median PA of 97.3% (IQR: 96.8%–97.7%) for 7-day horizon forecasts, whereas Reg and RF models with spatial smoothing had median PAs of 97.3% (IQR: 96.7%–97.7%) and 96.4% (IQR: 95.6%–96.7%), respectively. EpiNow2 with spatial smoothing obtained a median PA of 87.6% (IQR: 86.6%–88.9%).

As shown in **Table S4**, the XGB model with spatial smoothing outperformed all the other individual models in forecasting daily COVID-19 cases. In Scenario-1, RF with spatial smoothing achieved a median PA of 87.7% (IQR: 84.9%–89.7%) for 7-day horizon forecast, surpassing Reg and XGB models with spatial smoothing, which had median PAs of 86.7% (IQR: 83.1%–87.8%) and 86.8% (IQR: 83.4%–87.8%), respectively.

The forecasting performance of Reg, RF, and XGB models is evaluated through **Figures S15-S20** for R_t forecasts and **Figures S21-S26** for COVID-19 case forecasts, across different forecast horizons. For R_t forecasting, all Reg, RF, and XGB models consistently provided predictions, closely following the smoothed R_t estimates across different counties. For COVID-19 case forecasting in Scenario-1, XGB and RF with spatial smoothing outperformed the other models,

particularly in counties with complex transmission dynamics. In Scenario-2 case forecasting, XGB and RF without spatial smoothing achieved higher prediction accuracy compared to the other models.

4 Discussion

Accurate forecasting of R_t and COVID-19 case counts is critical for epidemic control and public health decision-making. Previous studies have demonstrated the importance of real-time R_t estimation in outbreak control, emphasizing its role in quantifying disease transmissibility and guiding policy responses [1,18]. Traditional statistical models, such as those used in Bayesian inference-based approaches like EpiNow2, have been widely employed to estimate R_t in real time [1,43]. This study integrates machine learning techniques and a probabilistic framework to address the challenges of forecasting disease transmission and cases. Our findings indicate that the ensemble-based approach outperformed the widely used EpiNow2 model across different forecasting horizons.

The practical implications of our forecasting framework are significant. Accurate R_t predictions enable early detection of outbreaks, allowing for timely interventions such as targeted lockdowns or vaccination campaigns. Similarly, reliable case forecasts inform resource allocation, ensuring that medical supplies, hospital beds, and testing infrastructure are optimally distributed. The ability to evaluate policy effectiveness through changes in R_t and forecasted trends supports evidence-based decision-making. Although our analysis focused on COVID-19 transmission across counties in SC, the proposed ensemble framework is generalizable to other infectious diseases and spatial scales. With appropriate input data, it can be applied to finer geographic

resolutions (e.g., ZIP code or census tract levels) or to different pathogens such as influenza, or RSV to support region-specific public health decision-making.

While the proposed forecasting framework demonstrated improved accuracy compared to EpiNow2, several limitations must be acknowledged. Previous studies have shown that geographically refined predictions can better inform targeted responses, including mobile health clinic (MHC) deployments [44]. More detailed spatial forecasts would allow public health officials to allocate resources more effectively, particularly in underserved communities where healthcare access is limited. Additionally, real-time forecasting remains challenging due to potential delays in obtaining case data from state health departments. In practice, alternative data sources, such as electronic health records (EHR) from health systems, may be necessary to supplement official reports. Previous research has demonstrated promise in leveraging alternative data sources for predictive modeling, but adapting models for real-time implementation in such settings requires further validation [44]. Furthermore, forecast accuracy declines as the prediction horizon increases, a common issue in time series forecasting.

Future research could focus on integrating Bayesian deep learning techniques with real-time data streams, such as mobility, genomic surveillance, wastewater data, and electronic health records, to improve forecast accuracy and uncertainty quantification.

Conclusion

In this study, we developed a forecasting framework to predict the effective reproductive number (R_t) and COVID-19 daily case counts using a combination of machine learning models and probabilistic approaches. By leveraging Bayesian estimates, spatial smoothing, and recursive forecasting, our approach enhances the accuracy of epidemic predictions. The ensemble-based

methodology further refines forecast by optimally combining individual model outputs using PA as a weighting criterion. Our results demonstrate that incorporating spatially adjusted estimates improves forecast performance across different pandemic waves and forecasting horizons. Future work should explore adaptive modeling techniques that integrate real-time mobility data and genomic surveillance to refine predictions further. Additionally, expanding the framework to accommodate multiple infectious diseases could enhance its applicability for broader epidemiological forecasting. Overall, this study contributes to advancing epidemic forecasting by integrating machine learning and probabilistic models in a flexible and scalable framework. These findings can support data-driven policy interventions, improve epidemic preparedness, and facilitate effective public health responses in future outbreaks.

Ethics statement

Ethical review for this study was obtained by the Institutional Review Board of Clemson University (2020-0150). No consent was needed for this study.

CRediT authorship contribution statement

LR, CM, SH conceptualized and developed the methodology for this study. SH conducted formal analysis, visualization, validation, data curation, and drafted the original manuscript with contributions from LR. LR, CM, and RG provided critical feedback and guidance throughout the research process. All authors contributed to interpretation of results. LR and CM oversaw project administration and supervision. LR acquired the funding for this study. All authors participated in reviewing and editing the manuscript.

Declaration of competing interest

SH, RG, NM, VG, TA, LR, and CM received salary support from the National Library of Medicine of the National Institutes of Health (R01LM014193). SH, TA, LR and CM received salary support from the Center for Forecasting and Outbreak Analytics of the Centers for Disease Control and Prevention (NU38FT000011). The authors declare no other potential conflicts of interest.

Acknowledgements

This project has been funded by the National Library of Medicine of the National Institutes of Health (NIH) under award number R01LM014193 and the Center for Forecasting and Outbreak Analytics of the Centers for Disease Control and Prevention (CDC) under award number NU38FT000011. The content and decision to publish is solely based on the authors of this study and does not necessarily represent the official views of the NIH or CDC. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of this manuscript.

Data and code availability

All code and public data used for analyses and figure generations are available at https://github.com/mdsakhh/Machine-Learning-and-Probabilistic-Approach-for-Rt-and-COVID-19-Case-Forecasting and from the corresponding author upon request.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT-4o and GPT-5 to assist with writing and grammar checks. After using these tools, the authors reviewed and edited the content as needed.

References

1. Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. Practical considerations for measuring the effective reproductive number, Rt. PLOS Comput Biol. 2020 Dec 10;16(12):e1008409.

2. Huisman JS, Scire J, Angst DC, Li J, Neher RA, Maathuis MH, et al. Estimation and worldwide monitoring of the effective reproductive number of SARS-CoV-2. Davenport MP, editor. eLife. 2022 Aug 8;11:e71345.

3. Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control. Oxford, New York: Oxford University Press; 1992. 766 p.

4. Nishiura H, Chowell G. The Effective Reproduction Number as a Prelude to Statistical Estimation of Time-Dependent Epidemic Trends. In: Chowell G, Hyman JM, Bettencourt LMA, Castillo-Chavez C, editors. Mathematical and Statistical Estimation Approaches in Epidemiology [Internet]. Dordrecht: Springer Netherlands; 2009 [cited 2023 Dec 19]. p. 103–21. Available from: http://link.springer.com/10.1007/978-90-481-2313-1_5

5. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178(9):1505–12.

6. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020 Aug;584(7820):257–61.

7. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PLOS ONE. 2007 Aug 22;2(8):e758.

8. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis. 2020 May;20(5):553–8.

9. Banholzer N, Feuerriegel S, Vach W. Estimating and explaining cross-country variation in the effectiveness of non-pharmaceutical interventions during COVID-19. Sci Rep. 2022 May 9;12(1):7526.

10. Lison A, Persson J, Banholzer N, Feuerriegel S. Estimating the effect of mobility on SARS-CoV-2 transmission during the first and second wave of the COVID-19 epidemic, Switzerland, March to December 2020. Eurosurveillance. 2022 Mar 10;27(10):2100374.

11. Abbott S, Hellewell J, Sherratt K, Gostic K, Hickson J, Badr HS, et al. EpiNow2: estimate real-time case counts and time-varying epidemiological parameters. R Package Version 01 0 [Internet]. 2020 [cited 2024 Feb 13]; Available from: https://scholar.google.com/scholar?cluster=13080972427650344 06&hl=en&oi=scholarr

12. Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. Am J Epidemiol. 2004 Sep 15;160(6):509–16.

13. Meyer S, Held L, Höhle M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. J Stat Softw [Internet]. 2017 [cited 2024 Mar 6];77(11). Available from: http://arxiv.org/abs/1411.0416

14. Bettencourt LM, Ribeiro RM. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. PloS One. 2008;3(5):e2185.

15. Cauchemez S, Boëlle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, et al. Real-time estimates in early detection of SARS. Emerg Infect Dis. 2006;12(1):110.

16. Thompson RN, Stockwin JE, van Gaalen RD, Polonsky JA, Kamvar ZN, Demarsh PA, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics. 2019 Dec 1;29:100356.

17. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci. 2006 Nov 28;274(1609):599–604.

18. Parag KV. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. PLOS Comput Biol. 2021 Sep 7;17(9):e1009347.

19. Bokányi E, Vizi Z, Koltai J, Röst G, Karsai M. Real-time estimation of the effective reproduction number of COVID-19 from behavioral data. Sci Rep. 2023 Dec 5;13(1):21452.

20. Nash RK, Nouvellet P, Cori A. Real-time estimation of the epidemic reproduction number: Scoping review of the applications and challenges. PLOS Digit Health. 2022 Jun 27;1(6):e0000052.

21. Won YS, Son WS, Choi S, Kim JH. Estimating the instantaneous reproduction number (Rt) by using particle filter. Infect Dis Model. 2023 Dec 1;8(4):1002–14.

22. Mellor J, Tang ML, Jones O, Ward T, Riley S, Deeny SR. Forecasting COVID-19, influenza, and RSV hospitalizations over winter 2023–4 in England. Int J Epidemiol. 2025 Jun 1;54(3):dyaf066.

23. Dixon S, Keshavamurthy R, Farber DH, Stevens A, Pazdernik KT, Charles LE. A Comparison of Infectious Disease Forecasting Methods across Locations, Diseases, and Time. Pathogens. 2022 Jan 29;11(2):185.

24. Tang CY, Gao C, Prasai K, Li T, Dash S, McElroy JA, et al. Prediction models for COVID-19 disease outcomes. Emerg Microbes Infect. 13(1):2361791.

25. Gibson GC, Moran KR, Reich NG, Osthus D. Improving probabilistic infectious disease forecasting through coherence. PLOS Comput Biol. 2021 Jan 6;17(1):e1007623.

26. Lutz CS, Huynh MP, Schroeder M, Anyatonwu S, Dahlgren FS, Danyluk G, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. BMC Public Health. 2019 Dec 10;19(1):1659.

27. Gao Y, Cai GY, Fang W, Li HY, Wang SY, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun. 2020 Oct 6;11(1):5033.

28. Mei X, Lee HC, Diao K yue, Huang M, Lin B, Liu C, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med. 2020 Aug;26(8):1224–8.

29. Wynants L, Calster BV, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020 Apr 7;369:m1328.

30. Arik S, Li CL, Yoon J, Sinha R, Epshteyn A, Le L, et al. Interpretable Sequence Learning for Covid-19 Forecasting. In: Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2020 [cited 2023 Nov 7]. p. 18807–18. Available from: https://proceedings.neurips.cc/paper/2020/hash/d9dbc51dc534921589adf460c85cd824-Abstract.html

31. Fife DA, D'Onofrio J. Common, uncommon, and novel applications of random forest in psychological research. Behav Res Methods. 2023 Aug 1;55(5):2447–66.

32. King C, Strumpf E. Applying random forest in a health administrative data context: a conceptual guide. Health Serv Outcomes Res Methodol. 2022 Mar 1;22(1):96–117.

33. Oidtman RJ, Omodei E, Kraemer MUG, Castañeda-Orjuela CA, Cruz-Rivera E, Misnaza-Castrillón S, et al. Trade-offs between individual and ensemble forecasts of an emerging infectious disease. Nat Commun. 2021;12(1).

34. Hossain MS, Goyal R, Martin NK, DeGruttola V, Chowdhury MM, McMahan C, et al. A flexible framework for local-level estimation of the effective reproductive number in geographic regions with sparse data. BMC Med Res Methodol. 2025 Mar 18;25(1):73.

35. The New York Times. Coronavirus (Covid-19) Data in the United States [Internet]. 2021 [cited 2023 Oct 6]. Available from: https://github.com/nytimes/covid-19-data

36. U.S. Census Bureau. Census Data [Internet]. 2020. Available from: https://data.census.gov/profile

37. Agency for Toxic Substances and Disease Registry. CDC/ATSDR SVI: Data and Documentation Download [Internet]. 2022. Available from: https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html

38. Grant R, Charmet T, Schaeffer L, Galmiche S, Madec Y, Platen CV, et al. Impact of SARS-CoV-2 Delta variant on incubation, transmission settings and vaccine effectiveness: Results from a nationwide case-control study in France. Lancet Reg Health – Eur [Internet]. 2022 Feb 1 [cited 2024 May 10];13. Available from: https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(21)00264-7/fulltext

39. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020 May 5;172(9):577–82.

40. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2020 Apr;93:284–6.

41. Harris JE. Timely epidemic monitoring in the presence of reporting delays: anticipating the COVID-19 surge in New York City, September 2020. BMC Public Health. 2022 May 2;22(1):871.

42. Hossain MS, Goyal R, Martin NK, DeGruttola V, Chowdhury MM, McMahan C, et al. A Flexible Framework for Local-Level Estimation of the Effective Reproductive Number in Geographic Regions with Sparse Data [Internet]. medRxiv; 2024 [cited 2024 Dec 12]. p. 2024.11.06.24316859. Available from: https://www.medrxiv.org/content/10.1101/2024.11.06.24316859v1

43. Abbott S, Lison A, Funk S. epinowcast: Flexible hierarchical nowcasting. Zenodo [Internet]. 2021 [cited 2024 Feb 13]; Available from: https://samabbott.co.uk/presentations/2023/royal-society-epinowcast.pdf

44. Gezer F, Howard KA, Litwin AH, Martin NK, Rennert L. Identification of factors associated with opioid-related and hepatitis C virus-related hospitalisations at the ZIP code area level in the USA: an ecological and modelling study. Lancet Public Health. 2024 Jun 1;9(6):e354–64.