JAMIA Open, 2024, **7(1)**, ooae014
https://doi.org/10.1093/jamiaopen/ooae014
Advance access publication 10 February 2024
Research and Applications

Research and Applications

Enhanced SARS-CoV-2 case prediction using public health data and machine learning models

Bradley S. Price, PhD[1,2,*], Maryam Khodaverdi, MS[2], Brian Hendricks, PhD[2,3], Gordon S. Smith, MD[2,3], Wes Kimble, MPA[2], Adam Halasz, PhD[4], Sara Guthrie, MS[5], Julia D. Fraustino, PhD[6], Sally L. Hodder, MD[2,7]

[1]Department of Management Information Systems, West Virginia University, Morgantown, WV 26505, United States, [2]West Virginia Clinical and Translational Science Institute, Morgantown, WV 26506, United States, [3]Department of Epidemiology and Biostatistics, West Virginia University, Morgantown, WV 26505, United States, [4]School of Mathematics and Data Science, West Virginia University, Morgantown, WV 26506, United States, [5]Department of Sociology and Anthropology, West Virginia University, Morgantown, WV 26505, United States, [6]Department of Strategic Communication, Reed College of Media, West Virginia University, Morgantown, WV 26505, United States, [7]Department of Medicine, West Virginia University, Morgantown, WV 26506, United States

*Corresponding author: Department of Management Information Systems, West Virginia University, 83 Beechurst Avenue Morgantown, WV 26505, United States (brad.price@mail.wvu.edu)

Dr B.S. Price and M. Khodaverdi contributed equally.

Abstract

Objectives: The goal of this study is to propose and test a scalable framework for machine learning (ML) algorithms to predict near-term severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cases by incorporating and evaluating the impact of real-time dynamic public health data.

Materials and Methods: Data used in this study include patient-level results, procurement, and location information of all SARS-CoV-2 tests reported in West Virginia as part of their mandatory reporting system from January 2021 to March 2022. We propose a method for incorporating and comparing widely available public health metrics inside of a ML framework, specifically a long-short-term memory network, to forecast SARS-CoV-2 cases across various feature sets.

Results: Our approach provides better prediction of localized case counts and indicates the impact of the dynamic elements of the pandemic on predictions, such as the influence of the mixture of viral variants in the population and variable testing and vaccination rates during various eras of the pandemic.

Discussion: Utilizing real-time public health metrics, including estimated R_t from multiple SARS-CoV-2 variants, vaccination rates, and testing information, provided a significant increase in the accuracy of the model during the Omicron and Delta period, thus providing more precise forecasting of daily case counts at the county level. This work provides insights on the influence of various features on predictive performance in rural and non-rural areas.

Conclusion: Our proposed framework incorporates available public health metrics with operational data on the impact of testing, vaccination, and current viral variant mixtures in the population to provide a foundation for combining dynamic public health metrics and ML models to deliver forecasting and insights in healthcare domains. It also shows the importance of developing and deploying ML frameworks in rural settings.

Lay Summary

This study aims to propose and test a scalable framework for machine learning (ML) algorithms to predict near-term severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cases by county by incorporating and evaluating the impact of real-time dynamic public health data. Data used in this study include patient-level results, procurement, and location information of all SARS-CoV-2 tests reported in West Virginia as part of their mandatory reporting system from January 2021 to March 2022. We propose a method for incorporating and comparing widely available public health metrics inside of a ML framework, specifically a long-short-term memory network, to forecast SARS-CoV-2 cases across various feature sets. Our approach provides better prediction of localized case counts, recommendation of locations of outbreaks, and indicates the impact of the dynamic elements of the pandemic on predictions, such as the influence of the mixture of viral variants in the population and variable testing and vaccination rates during various eras of the pandemic. Incorporating available public health metrics with operational data on the impact of testing, vaccination, and current viral variant mixtures in the population provides a foundation for combining dynamic public health metrics and ML models to deliver improved forecasting and insights in healthcare domains. This approach provides a model for utilizing ML to forecast, deploy, and understand the impact of public health data during coronavirus disease and other pandemics.

Key words: public health data; machine learning; SARS-CoV-2 prediction.

Received: September 20, 2023; **Revised:** January 29, 2024; **Editorial Decision:** February 5, 2024; **Accepted:** February 8, 2024

Background and significance

Forecasting the number of future severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cases and related hospitalizations has been a critical component in informing SARS-CoV-2 testing and pandemic response, especially regarding preparation for surges in hospitalization. Various epidemiological and mathematical models have been used to describe infectious disease transmission and predict the peak, duration, and magnitude of outbreaks.[1–3] Multiple methods have been used to forecast the number of SARS-CoV-2 cases from compartmentalized epidemiological models, such as the traditional suspected-infected-recovered (SIR) framework and various extensions that consider exposure, vaccination, and time frames of disease transmissibility (SEIR, SICCR, etc.).[4–6] Basic and instantaneous reproduction numbers (R_0 and R_t), the average secondary cases over time resulting from a primary case, are common public health metrics used to indicate the scale of epidemic spread within communities.[7–9] Metrics such as these may help researchers simulate infection spread through a population. However, the underlying assumptions can sometimes be overly simplified and may not reflect or make full use of the extensive and changing public health data that is available, ultimately resulting in unreliable estimates and forecasts.[10,11]

The gap between the mathematical frameworks and the data available to form actionable insights is evident in many of the metrics used during the coronavirus disease 2019 (COVID-19) pandemic response. For instance, the serial interval, which describes the time a secondary case will be detected after exposure to a primary case, is used in the estimation of R_t. However, R_t does not consider human test-seeking behaviors, an important determinant of case counts in public health data. Furthermore, many suggested models struggle to account for the variability in disease testing that results from changes in public policy and public sentiment or fatigue about the disease in question. Other issues that arise include enhanced transmissibility of emerging variants and potential for reinfection. Finally, assumptions regarding the ability of vaccines to prevent infection may not be directly built into the model, making forecasting cases problematic. All these issues must be considered for the optimization of predictive models.

Machine learning (ML) and artificial neural networks (ANNs) provide novel tools with which to address the COVID-19 pandemic.[12–15] Recurrent neural networks (RNNs) and networks that utilize long-short-term memory (LSTM) frameworks have shown superior performance for forecasting the number of infections compared to epidemiological models such as SIR and SEIR, or other non-ANN/statistical forecasting approaches such as ARIMA, GARCH, and PROPHET models.[16–18] Specifically, LSTM based models, both with and without spatiotemporal information, have also been used in predicting positive cases and deaths related to COVID-19.[18–22] While many of these models have shown merit at different points during the pandemic, the dynamic changes in the infectious agent, the methods by which data are recorded, and improved clinical care have resulted in model performance degrading over time. These issues create concern around the scalability of the ML techniques to correctly adjust to the changing dynamics of the science and the response.[23,24]

Objective

This study aims to propose and test a scalable framework for ML algorithms to predict near-term SARS-CoV-2 cases by incorporating real-time public health data. We also assess the impact of including various public health data and its changes over time on the performance and accuracy of the predictions and the resulting policy recommendation stemming from the predictions.

Methods

This retrospective study, conducted using data provided through a partnership with the West Virginia (WV) Department of Health and Human Resources (WVDHHR), received approval from the West Virginia University Institutional Review Board (IRB 2003952013 and 2011159080).

Data used in this study include patient-level results of all SARS-CoV-2 tests (both positive and negative) reported to WVDHHR as part of their mandatory reporting system from January 2021 to March 2022. This testing information also contains unique patient identifiers, test procurement data, patient zip code, and testing site location (including county). Due to many patients having multiple tests, only the first positive test for each individual in a 90-day window is used before July 2021, with the window being decreased to 60 days after July 2021 to account for enhanced reinfection potential with more recent variants (eg, Omicron).[25–27] For this study, all SARS-CoV-2 tests (both positive and negative) are aggregated to produce a data set that demonstrates the numbers of individual SARS-CoV-2 cases and testing rates by county and day per 10 000 residents during the period of interest. These data are combined with the total daily cumulative proportion of county residents who have received SARS-COV-2 vaccination. The 7-day uptake of both SARS-CoV-2 testing and vaccination per 10 000 residents is used as a proxy for community concern regarding COVID-19. By utilizing both metrics, weekly changes in testing and vaccination behavior can be calculated and analyzed. The Supplementary Material associated with this material provides examples of the testing and number of positive tests in various locations that were associated with this study. Finally, we calculate daily R_t values based on the serial intervals of Wuhan, Delta, and Omicron variants using the EpiEstm package in R.[28] The details for each serial interval are available in the GitHub repository associated with this article. Each of the 3 features (ie, R_t, vaccination, and provided tests) were utilized as inputs in candidate models (Figure 1). Other temporal information that may influence testing, including binary variables that indicate weekends or holidays, the number of days passed from the last major holiday, the number of days until the next major holiday, and county population, were also added to the input data.

Specifically, in this study, a multi-layer deep LSTM network was designed by stacking LSTM units trained by sliding window selections to forecast next week's cases. The 7-day sliding window removes the previous 7 days of data to provide a forecast for the day of interest, addressing any lags that may occur in data collection. Figure 2 shows the network structure and how the sliding input utilizes the input data from the sliding window. It consists of an input layer, followed by stacked LSTM with dropout layers and a fully connected dense layer. The fully connected dense layer was

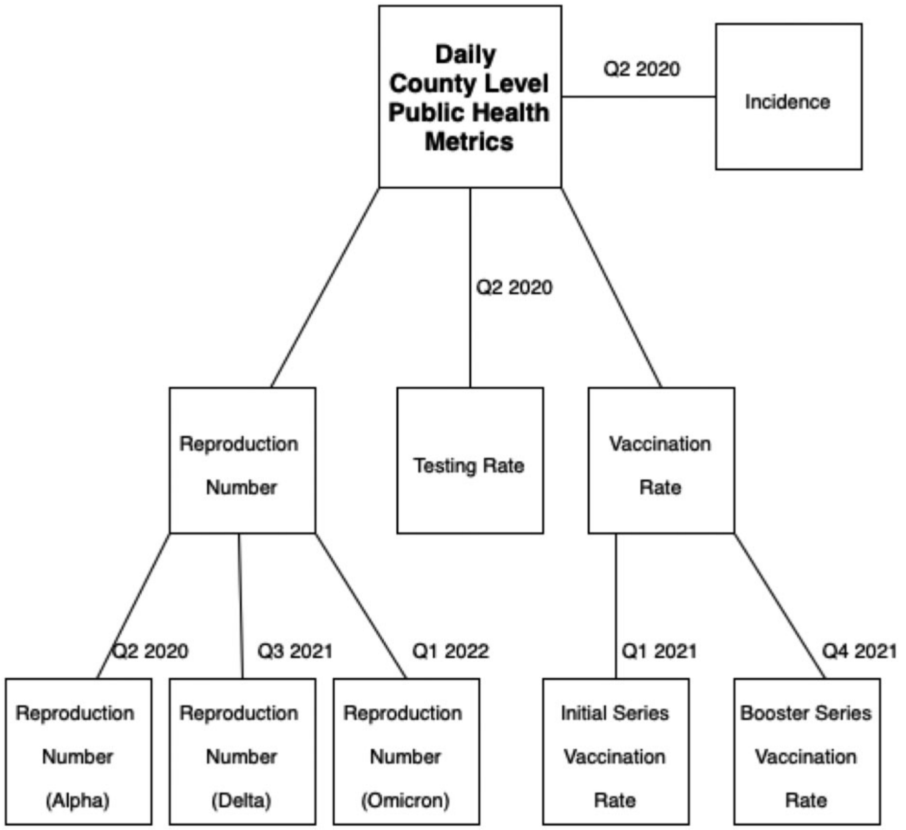

Daily County Level Public Health Metrics
Q2 2020
Incidence
Q2 2020
Reproduction Number
Testing Rate
Vaccination Rate
Q2 2020
Q3 2021
Q1 2022
Q1 2021
Q4 2021
Reproduction Number (Alpha)
Reproduction Number (Delta)
Reproduction Number (Omicron)
Initial Series Vaccination Rate
Booster Series Vaccination Rate

Figure 1. Structure of available public health data used for the framework with the quarter or of introduction indicated at the stem of the tree. Note other public health metrics could be used in addition to those listed.

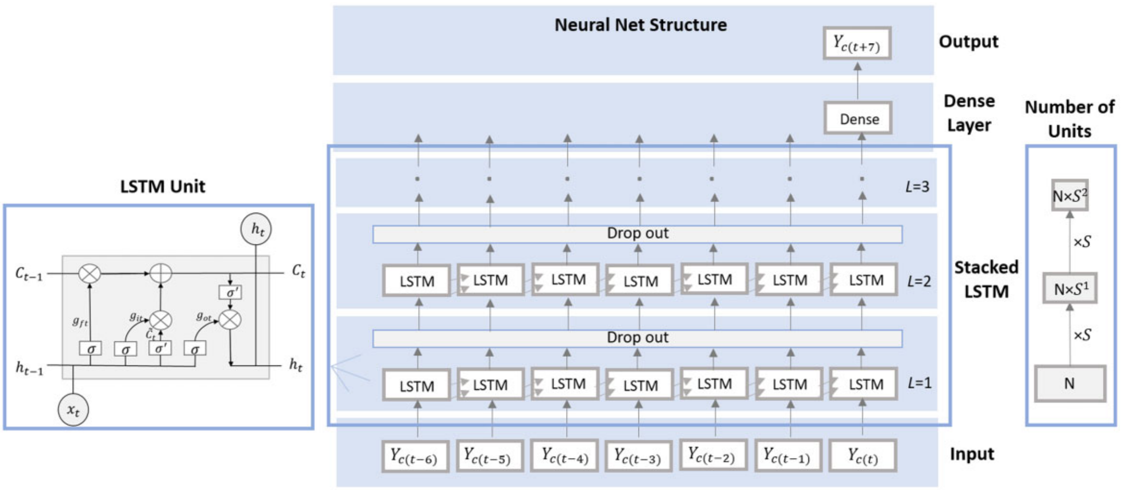
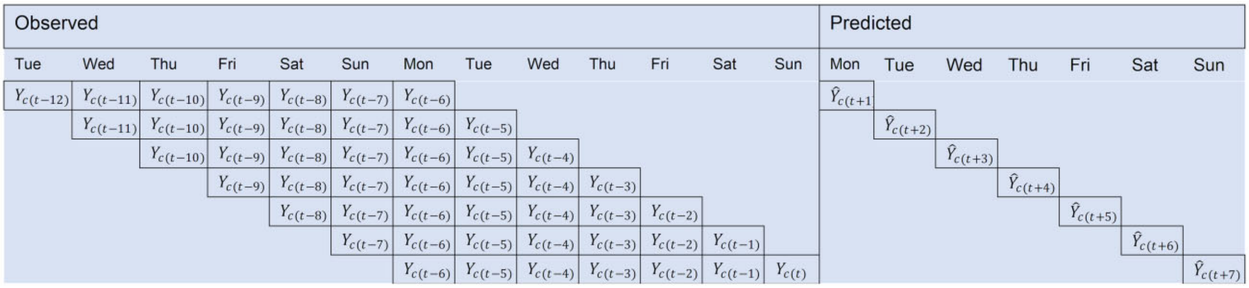


Figure 2. Proposed stacked LSTM network (middle), hidden unite sizes (right), and memory unite (left). Abbreviation: LSTM, long short-term memory.

added after stacked LSTM layers to interpret the output values with a linear activation function.

LSTM networks are RNNs, which have been widely used to solve multivariate time series forecasting problems due to possessing the ability to weight both long- and short-term trends (memories) appropriately.[29–31] Recursive connections within RNN allow previous information to persist inside of the network. In addition, long-term memories in LSTM networks allow the network to adjust to time series data appropriately and overcome the vanishing gradient problem of RNN, which does not let weights inside the network update appropriately.[32] As the shallow LSTM architecture might not be able to represent the complex features of sequential data efficiently, particularly when attempting to process highly non-linear and multivariate time series, we designed a deep LSTM including stacked LSTM layers one above another, which can circumvent the limitations of the conventional shallow LSTM model.[33] We control the number of LSTM units in each layer by utilizing an encoder structure where the number of LSTM units in each layer is defined by NS^{L-1}



Observed													Predicted						
Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
$Y_{c(t-12)}$	$Y_{c(t-11)}$	$Y_{c(t-10)}$	$Y_{c(t-9)}$	$Y_{c(t-8)}$	$Y_{c(t-7)}$	$Y_{c(t-6)}$							$\hat{Y}_{c(t+1)}$						
	$Y_{c(t-11)}$	$Y_{c(t-10)}$	$Y_{c(t-9)}$	$Y_{c(t-8)}$	$Y_{c(t-7)}$	$Y_{c(t-6)}$	$Y_{c(t-5)}$							$\hat{Y}_{c(t+2)}$					
		$Y_{c(t-10)}$	$Y_{c(t-9)}$	$Y_{c(t-8)}$	$Y_{c(t-7)}$	$Y_{c(t-6)}$	$Y_{c(t-5)}$	$Y_{c(t-4)}$							$\hat{Y}_{c(t+3)}$				
			$Y_{c(t-9)}$	$Y_{c(t-8)}$	$Y_{c(t-7)}$	$Y_{c(t-6)}$	$Y_{c(t-5)}$	$Y_{c(t-4)}$	$Y_{c(t-3)}$							$\hat{Y}_{c(t+4)}$			
				$Y_{c(t-8)}$	$Y_{c(t-7)}$	$Y_{c(t-6)}$	$Y_{c(t-5)}$	$Y_{c(t-4)}$	$Y_{c(t-3)}$	$Y_{c(t-2)}$							$\hat{Y}_{c(t+5)}$		
					$Y_{c(t-7)}$	$Y_{c(t-6)}$	$Y_{c(t-5)}$	$Y_{c(t-4)}$	$Y_{c(t-3)}$	$Y_{c(t-2)}$	$Y_{c(t-1)}$							$\hat{Y}_{c(t+6)}$	
						$Y_{c(t-6)}$	$Y_{c(t-5)}$	$Y_{c(t-4)}$	$Y_{c(t-3)}$	$Y_{c(t-2)}$	$Y_{c(t-1)}$	$Y_{c(t)}$							$\hat{Y}_{c(t+7)}$

Figure 3. Predicting next week's daily number of cases with 7-day moving time windows.

where N is the original number of LSTM units in the network, S is a user-defined parameter between 0 and 1, and L is the layer of the network. Additional dropout layers are added to remove redundant information by ignoring randomly selected neurons during the training process.[34–36]

The LSTM memory units control the flow of information in the recurrent hidden layer (Figure 2). The memory units contain memory cells storing the temporal state of the network in addition to gates. Inside each memory unit, there are 3 gates (g): forget gate (g_{ft}), input gate (g_{it}), and output gate (g_{ot}), which control the cell state (C_t).

Information from the memory cell is propagated to the output with the activation of the output gate. The forget gate removes past memory cells' status, while the input gate accumulates information from the previous memory cells.[29] Summaries of relevant historical behavior are all collected and then passed to future cells. Define t as the time step, w as the weight variables, and b as the bias variables. Recurrent activation and activation are shown by σ and σ', which are defined as the sigmoid and hyperbolic tangent functions, respectively.

The updating equations at time point t for a given LSTM cell state C_t, and output h_t is calculated with the following formulas given input data, x_t:

$$W_g = [W_f,\ W_i,\ W_o],\ b_g = [b_f,\ b_i,\ b_o],$$

$$[g_{ft},\ g_{it},\ g_{ot}] = \sigma(W_g.[h_{t-1}, x_t] + b_g)$$

$$C_t = g_{ft} \times C_{t-1} + g_{it} \times \sigma'^{(W_c.[h_{t-1},x_t]+b_c)}$$

$$h_t = g_{ot} \times \sigma' C_t.$$

A sliding window approach for inputs is also used to represent the epidemiological relationship between the current number of cases and cases in the past 7 days. The LSTM framework was used as we anticipated complex temporal relationships between covariates and the associations defined by the cells. The designed network considered the past 7 days for each county as input and subsequently returned as an output a prediction of the number of positive cases for the county. The input data were defined by matrix $'Y_{c,t} = [Y_{c,t-6},\ Y_{c,t-5}, \ldots, Y_{c,t}]$, where $Y_{c,t}$ is the vector of inputs for county c at time step t. The predicted output is then, $\hat{Y}_{c,t+7}$, which is the predicted daily number of cases for county c at time step $t+7$, is our output. As is typical in reporting the number of positive cases from the procurement date, there is a reporting lag which previous studies have estimated as around 3 days (as of November 2021). To adjust for this lag from procurement to the result, positive cases are imputed from the sequence of the previous 3 days. Figure 3 shows an example of how observed data is used in our algorithm to generate a given prediction for a 7-day period.

As previously mentioned, the LSTM utilizes input data provided by utilizing the reproduction number (R_t) of 3 SARS-CoV-2 variants throughout this study. For the estimation of the Alpha, Delta, and Omicron variant R_t, we utilize the serial interval estimation technique proposed by Price et al.[18] By creating an independent variable for each rate of transmission of the respective variants, the LSTM, through the weighting and connected network, can better identify the importance of the rate of transmission associated with each variant. Utilizing this with frameworks for interactivity in the LSTM, we can mix the spread parameters through the input layers and interactions that occur in the network. The dropout layers also allow the mixture of the variants to be prioritized with various weights at various times.

Our network was trained using Adaptive Moment Estimation (ADAM) to minimize mean absolute error (MAE). Bayesian optimization was utilized to tune hyper-parameters, which include the learning rate, dropout rate, number of batches, number of epochs, and number of LSTM hidden units for each LSTM layer (which requires selecting N and S). Full details of the implementation are available in a GitHub repository associated with this article.

The input dataset was divided into a training set and a test set by using 80% of our available data as the training set and 20% of the most recent data as the validation set.[37] Shapley Additive Explanations (SHAP) values on the model with all variables included are used to identify and quantify the importance of the variable/feature and understand each variable's impact on the larger model.[38–40] SHAP values present an interpretable tool for understanding the importance of variables in predictive models such as LSTM and define the contribution of each variable to the prediction.[41,42] Positive SHAP values indicate a variable's positive influence or effect of a variable, that is an increase in that variable will lead to an increase in the predicted number of cases. Similarly, a negative SHAP value leads to a negative association between a covariate and the predicted number of cases. For comparison purposes, we explore the changing variable importance through SHAP values for various periods during the pandemic. Finally, to further show the importance of the variables used as inputs in the model, we compare 6 candidate models using various variable/feature combinations (see Table 1) utilizing the LSTM described previously. Each of the 6 models is designed to evaluate the importance and impact of a specific feature (or set of features) to near-term prediction of SARS-CoV-2 cases and the ability to identify locations

Table 1. Feature set of models of interest.

	Feature used in model (Yes/No) (Omicron time period)								
Model name	R_t alpha	R_t delta	R_t omicron	7-Day vaccine uptake	Cumulative vaccine rate (2 doses)	Number of tests given	Daily incidence count	90-Day rolling incidence count	240-Day rolling incidence count
Full Model	Y	Y	Y	Y	Y	Y	Y	N	N
Alpha R_t Only	Y	N	N	Y	Y	Y	Y	N	N
All R_t + tests	Y	Y	Y	N	N	Y	N	N	N
All R_t + Incidence	Y	Y	Y	N	N	N	Y	N	N
Full Model + 90-day incidence	Y	Y	Y	Y	Y	Y	Y	Y	N
Full Model + 240-day incidence	Y	Y	Y	Y	Y	Y	Y	N	Y

of case increases. Specifically, we compare methods utilizing more information for predicting case counts and recommending locations for interventions to a method using Alpha R_t[18] with the inclusion of information on vaccination, SARS-CoV-2 testing, and the inclusion of rolling incidence over various time frames. Note the Alpha R_t Only model is an updated version of the model proposed by Price et al and, under certain conditions of hyperparameters, would be equivalent, with the additional ability to utilize updated vaccination information. We also include features of rolling incidence to indicate the number of recovered patients in each locality; this information would be available to state-of-the-art epidemiological models. Models are evaluated on prediction accuracy using multiple metrics, including root mean squared error (RMSE), MAE, and mean absolute percentage error (MAPE) on the difference between forecasted and actual cases for each week during various periods of interest.

Let y_{tc} and $\hat{y}_{tc}$ be the actual and forecasted values for a 7-day period t at location c. We define the metrics RMSE, MAE, and MAPE as follows:

$$\text{RMSE} = \left((\text{DC})^{-1}\sum_{t=1}^{D}\sum_{c=1}^{C}(y_{tc} - \hat{y}_{tc})^2\right)^{1/2}$$

$$\text{MAE} = (\text{DC})^{-1}\sum_{t=1}^{D}\sum_{c=1}^{C}|y_{tc} - \hat{y}_{tc}|$$

$$\text{MAPE} = (\text{DC})^{-1}\sum_{t=1}^{D}\sum_{c=1}^{C}\frac{|y_{tc} - \hat{y}_{tc}|}{y_{tc}}$$

Note this metrics could also be calculated directly for various weeks by adjusting the calculations to only sum over the locations rather than the time periods.

Given the variation in numbers of SARS-CoV-2 tests and cases during this period and the small population sizes in some WV counties, comparisons of model prediction ability will only be assessed for situations when a county report has >10 cases over a 7-day period over 2 study phases from March 2020 to April 2022. All computation was performed on a single Nvidia GPU and implemented in Python 3.8.

To assess the impact of these approaches on testing location recommendations, we utilize binary discount cumulative gain (BDCG), like the approach used by Price et al,[18] to recommend the top 10 counties for enhanced SARS-CoV-2 testing for a given period of interest. To keep from biasing the evaluation toward rural areas with a low incidence, we only consider those with $y_{c,t} > 10$. Define S_t to be the set of indices, the largest 10 values of $\frac{y_{c,t+1}}{y_{c,t}}$ for a given time point. We define the BDCG of a set of rankings at time point t as:

$$\sum_{i=1}^{q}\frac{I(i \in S_t)}{\ln(i+1)},$$

where $I(i \in S_t)$ is an indicator of a correct identification of a top 10 ranking in the actual percentage increases, and q is the number of rankings used in the calculation. For example, if $q = 10$, then BDCG$_t$ would only evaluate the top 10 rankings, in our setting, this would be the top 10 counties, returned by a method. One may view BDCG as a weighted identifier to measure the quality of the rankings for purposes of identifying case increases (or spikes) of the top q recommendations.

Two time periods were of interest during deployment: Delta and Omicron waves. The first analysis considers models trained on data before Q1 (January-March) 2022 and then evaluates and compares the results using data from Q1 2022. Specifically, this analysis focuses on the dynamics of the feature set performance during the Omicron period in WV. The second analysis is an experiment to validate and extend results from the first study and considers models trained on data before and including Q3 (July-September) 2021 and evaluates and compares the MSE of models using data just from Q4 (October-December) 2021, thereby providing a focus on models that are trained and evaluated when the Delta variant was the primary variant in WV. Note during the Delta evaluation and training period, we remove the R_t Omicron feature from the model as it would not have yet been known to mimic the deployment of the model. We focus on the accuracy of the models (RMSE, MAE, and MAPE) the quality of the recommendations (BDCG), and interpretations of the models (using SHAP values derived from the Full Model to assess the impact of each variable).

Results

During Q4 2021 (the Delta wave) and Q1 2022 (the Omicron wave), the Full model which incorporates R_t of multiple viral variants, vaccination, and testing information has consistently lower error rates as assessed by all metrics (RMSE, MAE, and MAPE) as compared to the other models (Table 2). During the Omicron period, the Full model predicts on average, 17.11 cases above or below the actual value as evaluated

Table 2. A comparison of multiple error rates over a set of models of interest based on various feature sets.

		Full Model	Alpha R_t Only	All R_t + Testing	All R_t + Incidence	Full Model + 90-day rolling incidence	Full Model + 240-day rolling incidence
Delta time period (Q4 2021)	RMSE	10.80	11.46	18.74	19.65	10.97	11.31
	RMSE >10	14.06	15.00	24.98	25.26	14.26	14.75
	MAE	5.83	5.93	7.27	7.91	5.80	8.23
	MAE >10	8.25	8.45	10.87	12.06	8.15	8.23
	MAPE >10	27.66%	28.07%	32.87%	36.60%	27.74%	27.85%
Omicron time period (Q1 2022)	RMSE	17.11	17.39	21.82	29.26	17.26	19.75
	RMSE >10	23.43	23.82	29.95	40.23	23.65	27.10
	MAE	8.05	8.06	9.54	12.71	8.05	8.77
	MAE >10	13.50	13.55	16.32	22.25	13.51	14.70
	MAPE >10	26.13%	25.98%	30.66%	48.27%	25.97%	26.39%

The models are evaluated on data during Q4 2021 and Q1 2022, while trained on all data from prior time periods. Model names and feature sets are referenced in Table 1.

by RMSE for all cases and 23.43 cases when the actual number of cases is larger than 10. In the case of MAE and MAPE, the Full model predicts on average above or below by 8.05 cases and by 26.13%, respectively. During the Delta period, a similar result is shown with RMSE for all cases and for RMSE when the number of actual cases was larger than 10, showing that the Full model predicts 10.80 and 14.06 cases of the observed number of cases, respectively. When considering MAE and MAPE, results show that the Full model predicts, on average, above or below by 5.83 cases and by 27.66%, respectively.

Table 2 provides a deeper comparison of all models aggregated over each of the 2 study periods for comparison. For instance, when directly comparing the Full model, which uses all available public health information during the Omicron period to the model that omits the variables containing the Omicron and Delta R_t (but includes information on the Alpha R_t Only), the RMSE is lowered by ∼1.5%. We also note that the models containing the full model, in addition to rolling incidence, have a lower MAE by 1.2% in Delta and a lower MAPE in Omicron, though this performance is not consistent across all metrics. While this does provide some level of understanding of how the models performed over each period, the aggregation leaves out details of how the models may change over each of the time periods. Figure 4 shows the BDCG, RMSE, and MAPE weekly comparisons for each of the 2 time periods, respectively, for all 55 counties with case counts >10. The results presented in Figure 4 show that the Alpha R_t Only and Full Model consistently have a lower RMSE and MAPE than the other 2 models at all time points but have similar RMSE and MAPE. The results also show the fluctuation in the results as time changes, and the dynamics on the ground influence the effectiveness of the models in practice. The rural and non-rural comparison presented in Figure 5 provides even further insight into the performance of the models during each of the periods. Specifically, we find that the Full Model has a lower RMSE for all non-rural counties during the Q1 2022 testing period (Omicron time period). The results also show that in the latter part of the Q4 2021 testing period, the Full Model has a lower RMSE for both rural and non-rural counties. We find that the models that showed performance improvements from the rolling incidence models are in the non-rural locations during both the Delta and Omicron testing periods.

During the Omicron testing period, the results show that the models with the feature sets of All R_t + Incidence or All R_t + Testing models have the largest BDCG. These results show in 8 out of the 11 weeks during the Omicron testing period the best-performing models do not have vaccination information as a variable. Conversely, during the Delta deployment period, we see that the Full Model has the largest or second largest BDCG in 4 of the 5 first weeks. From that point on All R_t + Incidence has the highest BDCG for the rest of the period.

Figure 6 shows the relevant SHAP values corresponding to the Omicron and Delta surges in WV, respectively. The SHAP values presented show that during the Omicron surge the R_t associated with the serial interval of the Omicron variant had the greatest positive association with case increases, followed by the 7-day vaccine up-take metric, while cumulative vaccination rate and the R_t associated with the Alpha variant serial interval is associated with lower case count forecasts. Similarly, for the Delta variant surge, 7-day uptake and the R_t associated with the delta variant serial interval are shown to be associated with increased case count forecasts, and cumulative vaccination rate is associated with lower case count forecasts.

Discussion

Our model utilizing real-time public health metrics including estimated R_t from multiple SARS-CoV-2 variants, vaccination, and testing information provided a significant decrease in the RMSE of the model during the Omicron and Delta period, thus providing a more precise forecasting of daily case counts at the county level. Our methodology is likely scalable beyond the COVID-19 pandemic and provides a framework for utilizing real-time public health data as it becomes available during future epidemics.

Using multiple R_t values in the model is novel and addresses the divergent behavior of different SARS-CoV-2 variants in the population. Due to the cost and logistics of conducting real-time SARS-CoV-2 genomic sequencing, the mixture of the variants at any given time is estimated. R_t is an estimated rate of transmission based on a serial interval. By utilizing this estimated or apparent R_t, a mixture of viral variants may be estimated in the model that best predicts the spread of cases. The serial interval changed as new SARS-CoV-2 variants with enhanced transmissibility entered the population. However, consideration must be given to 2 issues: (1) the actual serial interval of the variant may not be the appropriate interval for use in predicting future case

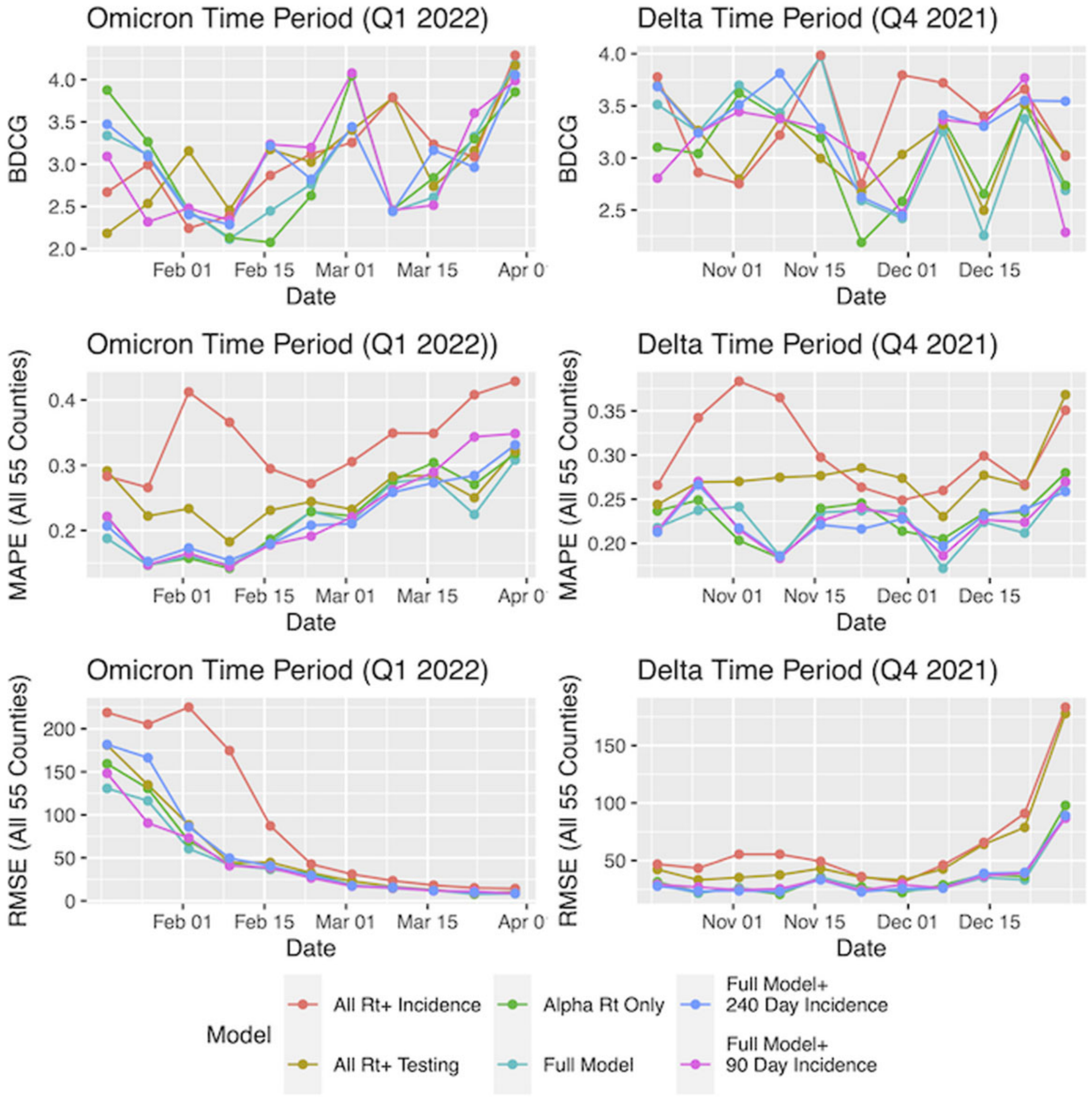


Figure 4. A comparison of weekly BDCG of top 10 Locations (top), MAPE (middle), and RMSE (bottom) of each of the 6 models in the 2 time periods of our study. Abbreviations: BDCG, binary discount cumulative gain; MAPE, mean absolute percentage error; RMSE, root mean squared error.

numbers using public health data which are based on community testing behaviors and result reporting; (2) even If the estimate of R_t is biased (eg, the serial interval used is not representative of the apparent serial interval), it is still valid to interpret R_t as a rate of transmission over a given time period, and our framework provides the ability to utilize multiple R_t values to understand better the number of cases generated from a single case given this bias. While there was little change in the error measures when we compared the model using just alpha to adding in other variants, this may not be the case for other emerging infections.

The value of the proposed framework that considers testing uptake, vaccination, and infection prevalence in the community is that it may adjust dynamically based on trends that occur in the population. For instance, results from Q1 2022 show an increase in forecasted case counts with the increase of R_t associated with the Omicron variant but lower-case counts associated with R_t associated with the earlier Alpha variant (Figure 6). Given that the Alpha variant was not detected as present in the population during this time, it presents an interesting feature of the methodology. This dynamic between 2 spread parameters could be indicative of co-occurring processes in the population. The first is simply that if the 2 R_t values are unrelated, that is increase the R_t of the Alpha variant without a corresponding increase in the R_t of the Omicron variant, it would suggest that the increases occur outside of the normal window of spread associated with the Omicron variant. A second interpretation is that there is an increase in both R_t values; however, the calculated serial interval for the Omicron variant is not optimal regarding predicting incidences; the apparent serial interval of Omicron can be approximated by utilizing multiple variables inside of this ML model. In either case, this shows how our proposed approach can adjust for the misalignment of real-time public health data with underlying assumptions of mathematical models. Furthermore, an open area of research is utilizing frameworks and approaches such as this to provide crude estimates of the mixture of the variants in the population, as we have previously discussed, care must be taken not to confound population behaviors.

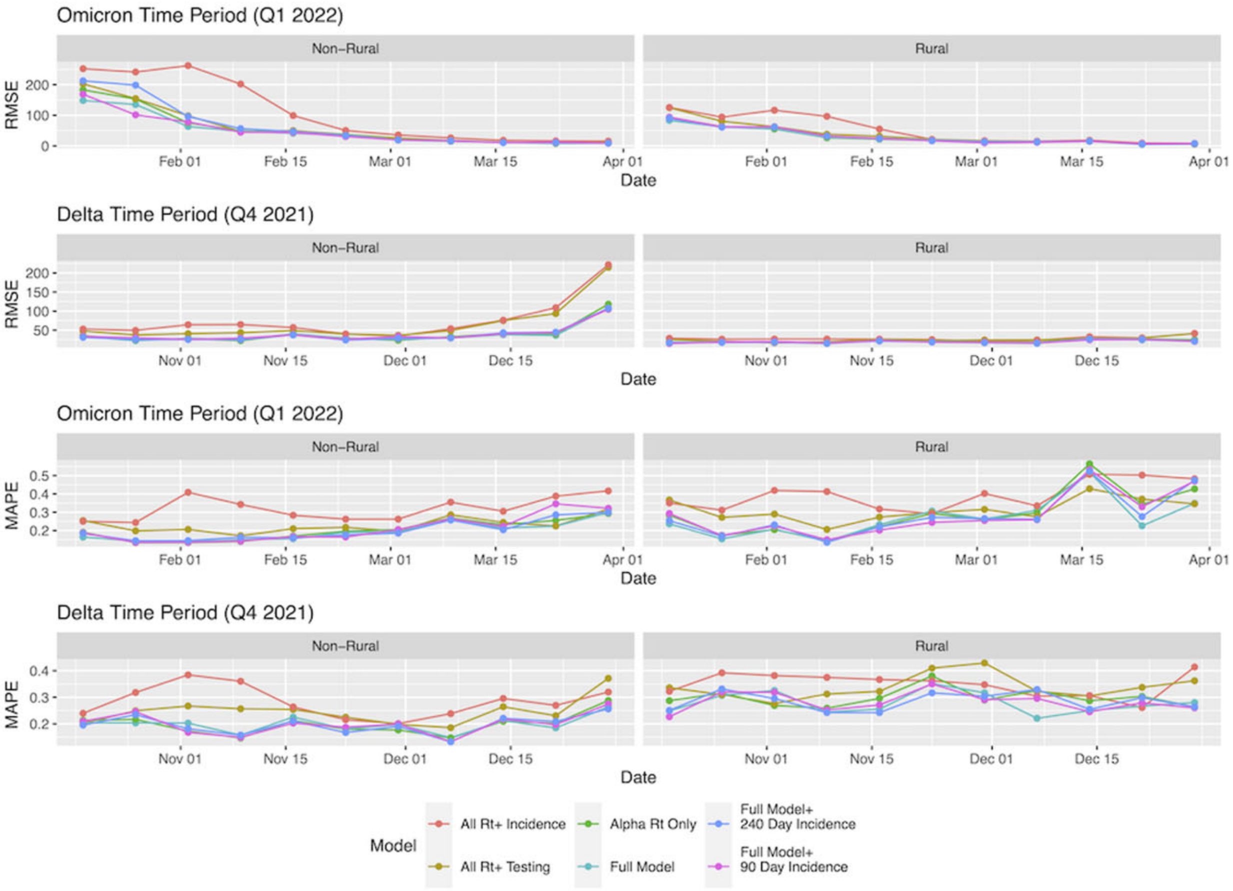


Figure 5. A comparison of weekly RMSE and MAPE of each of the 6 models stratified by rural and non-rural counties in the 2 time periods of our study. Abbreviations: MAPE, mean absolute percentage error; RMSE, root mean squared error.

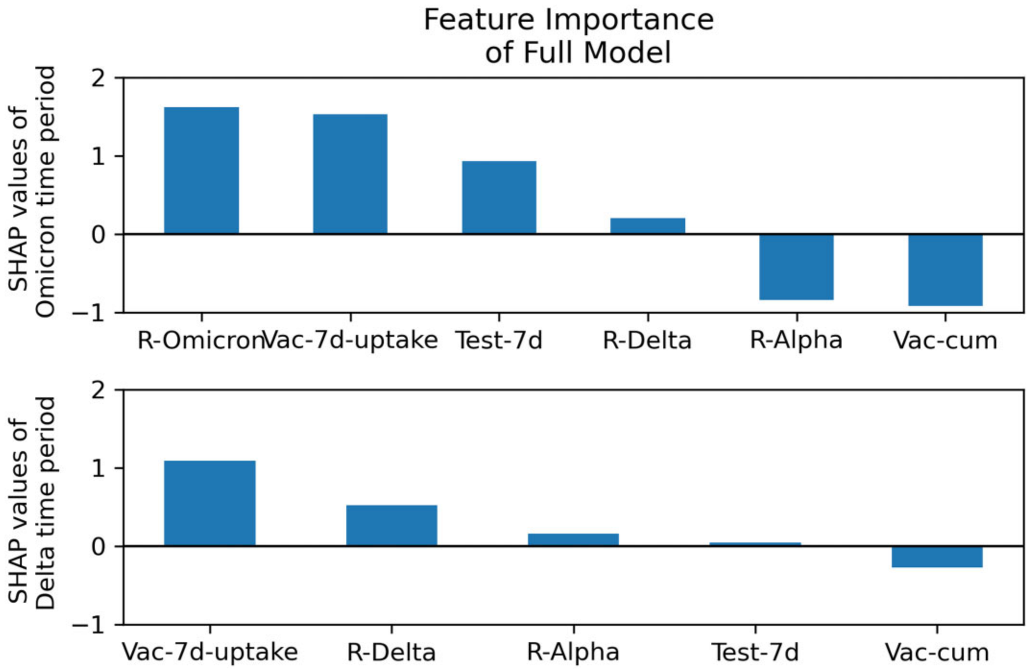


Figure 6. SHAP features importance values for models developed and deployed during the Omicron Era (Q1 2021) (top), based on models developed and deployed during the Delta era (Q3 2021) (bottom). Abbreviation: SHAP, Shapley Additive Explanations.

The proposed ML approach allows maximal use of all available data during the dynamic changes that occur in a public health crisis. As new data are introduced, we can directly calculate vaccination rates and variant-specific R_t values when necessary. Finally, we can account for and quantify the impact of testing and vaccination uptake variability and vaccination efficacy on subsequent case counts at specific times during the crisis while utilizing all historical data available. Again, we note that the 7-day vaccination and testing uptake metrics are used as a proxy for community concern regarding SARS-CoV-2 in the county.

The results show associated impacts that align with scientific knowledge about the impact of the variables. For example, the SHAP values show more testing is associated with higher case count forecasts, and higher cumulative vaccination is associated with lower case count forecasts for both Q3 2021 (Delta era) and Q1 2022 (Omicron era). The results presented in Table 2 and across Figures 4 and 5 further support the importance of using both testing and vaccination information when predicting case counts. The Full Model and Alpha R_t Only model, which uses all available information on testing and vaccination information, has lower RMSE in locations with both high and low (<10) daily case counts, suggesting relevance for rural areas where sparse population may be associated with low case counts. The ability of these models to provide accurate near-term predictions of SARS-CoV-2 cases, reflects the importance of changing clinical inputs with, the advantage of utilizing real-time public health data with an LSTM ML framework compared to traditional epidemiological approaches.

This study demonstrates insights into the behavior of ML models to predict case numbers that are so important for decisions regarding resource allocation in public health settings. If the purpose is to predict the number of cases across the state of WV, the Alpha R_t models and Full Model show superior performance, with each model performing better at various times and for rural and non-rural settings. We find that the Full Model and the All R_t + Incidence perform well at different times. Specifically, the model incorporating All R_t + Incidence performs better on BDCG, indicating that the vaccination information may no longer help identify locations where outbreaks are most likely to occur. We note that in the Omicron wave, vaccination information is only found in the best model used in 3 of the 11 weeks evaluated. From a ML perspective, this result helps alleviate concerns that these frameworks will always select models with the maximum number of available features. The performance of the R_t + Incidence model in predicting locations with case increases may be attributed to diminished vaccine effectiveness during the Omicron wave in WV. This approach was deployed during our NIH-funded RADX-Up project in WV, specifically during the Omicron surge to identify locations to approach for the deployment of expanded testing resources. Both recommended locations and forecasted cases were conserved when deploying these resources. When viewed in totality, these analyses illustrate the importance of utilizing all available public health information and real-time updating of the models, specifically implementing techniques such as online learning and other real-time implementation strategies for ML methods.[43]

A limitation of this work is that our approach utilizes cumulative vaccination and 7-day vaccination uptake as a proxy for community concern regarding the COVID-19 pandemic, but other metrics, such as second dose of vaccine boosters in eligible populations, as well as community reinfection rates, could also be important, especially since vaccination may lessen but not necessarily prevent infection. These metrics may also differ for various regions and should be considered in future research. Another limitation is that this study was only developed and deployed in West Virginia, though the approach we use is broadly applicable and could be deployed directly in other areas. Finally, change point detection is still an open problem in the ML community as the ability for models to update instantaneously in response to changes in underlying dynamics, such as human testing behaviors and viral transmissibility, is not yet available. Thus, there is a lag between the introduction of a variant and the resulting impact of the emergent variant in the model.

Conclusion

One of the key aspects of the public health response to any emerging outbreak is early detection through surveillance and then having accurate means of forecasting how it will impact the community. Outbreak analytics and disease forecasting, however, are only as reliable as the underlying data upon which forecasting models are based, and the data often changes rapidly over time. The disconnect between forecasting methods for SARS-CoV-2 cases and using the available real-time and changing public health metrics creates an operational gap when predicting case counts in pandemics. There is also a lack of data and studies that investigate, develop, and deploy ML methods in rural settings like that presented here. Novel methods such as ours are required to combine the useful information contained in widely available public health metrics with historical data trends. Our proposed framework provides a method to nimbly incorporate the available public health metrics with operational data and integrate insights on the impact of testing, vaccination, and current viral variant mixtures in the population to provide better near-term predictions of SARS-CoV-2 incidence. The framework also accounts for the dynamic changes in these features and the addition of new data as they become available. This also provides a model for utilizing interpretable ML techniques to forecast and understand the impact of public health data, including measuring metrics of emerging variants during COVID and other pandemics. Thus providing more precise forecasting of public health metrics that can be used to target disease control interventions and resource deployment more accurately, increasing their impact.

Author contributions

B.S.P. contributed to the methodological design, implementation, data analysis, figure creation, funding, interpretation, and writing. M.K. contributed to methodological design, implementation, data analysis, figure creation, and writing. B.H. contributed to methodological design, implementation, funding, and writing. G.S.S. contributed to methodological design, implementation, interpretation, and writing. W.K. contributed to methodological design, data collection, and writing. A.H. contributed to methodological design and writing. S.G. contributed to implementation, evaluation, and writing. J.D.F. contributed to the evaluation and writing. S.L. H. contributed to methodological design, implementation, funding, and writing.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by the National Institute of Health National Institute of Minority Health and Health Disparities RADx-Up 1U01MD017419-01, National Institute of General Medical Sciences RADx-Up 3U54GM104942-06S1, and National Institute of General Medical Sciences, 5U54GM104942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data used in this article is owned by a third party. To request access, contact the West Virginia Department of Health and Human Resources (https://dhhr.wv.gov/Pages/contact.aspx). Relevant code and data similar to what is used in this article can be found in the Github Repository: https://github.com/MKhodaverdi/COVID19-Enhanced-Case-Prediction/tree/main.

Reference

1. Padmanabhan R, Abed HS, Meskin N, Khattab T, Shraim M, Al-Hitmi MA. A review of mathematical model-based scenario analysis and interventions for COVID-19. *Comput Methods Programs Biomed*. 2021;209:106301. https://doi.org/10.1016/j.cmpb.2021.106301
2. Budd J, Miller BS, Manning EM, et al. Digital technologies in the public-health response to COVID-19. *Nat Med*. 2020;26(8):1183-1192. https://doi.org/10.1038/s41591-020-1011-4
3. Britton T, Ball F, Trapman P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science*. 2020;369(6505):846-849. https://doi.org/10.1126/science.abc6810
4. Xiang Y, Jia Y, Chen L, et al. COVID-19 epidemic prediction and the impact of public health interventions: a review of COVID-19 epidemic models. *Infect Dis Model*. 2021;6:324-342. https://doi.org/10.1016/j.idm.2021.01.001
5. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet*. 2020;395(10225):689-697. https://doi.org/10.1016/S0140-6736(20)30260-9
6. Carli R, Cavone G, Epicoco N, et al. Model predictive control to mitigate the COVID-19 outbreak in a multi-region scenario. *Annu Rev Control*. 2020;50:373-393. https://doi.org/10.1016/j.arcontrol.2020.09.005
7. Inglesby TV. Public health measures and the reproduction number of SARS-CoV-2. *JAMA*. 2020;323(21):2186-2187. https://doi.org/10.1001/jama.2020.7878
8. Van den Driessche P, Watmough J. Further notes on the basic reproduction number. In: Brauer F, van den Driessche P, Wu J, eds. *Mathematical Epidemiology*. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer; 2008:159-178. https://doi.org/10.1007/978-3-540-78911-6_6
9. Sera F, Armstrong B, Abbott S, et al.; CMMID COVID-19 Working Group. A cross-sectional analysis of meteorological factors and SARS-CoV-2 transmission in 409 cities across 26 countries. *Nat Commun*. 2021;12(1):5968. https://doi.org/10.1038/s41467-021-25914-8
10. Wang JIN. Mathematical models for COVID-19: applications, limitations, and potentials. *J Public Health Emerg*. 2020;4:9-9. https://doi.org/10.21037/jphe-2020-05
11. Fiscon G, Salvadore F, Guarrasi V, Garbuglia AR, Paci P. Assessing the impact of data-driven limitations on tracing and forecasting the outbreak dynamics of COVID-19. *Comput Biol Med*. 2021;135:104657. https://doi.org/10.1016/j.compbiomed.2021.104657
12. Dias SB, Hadjileontiadou SJ, Diniz J, et al. DeepLMS: a deep learning predictive model for supporting online learning in the Covid-19 era. *Sci Rep*. 2020;10(1):19888. https://doi.org/10.1038/s41598-020-76740-9
13. Subudhi S, Verma A, Patel AB, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *NPJ Digit Med*. 2021;4(1):87-87. https://doi.org/10.1038/s41746-021-00456-x
14. Shorten C, Khoshgoftaar TM, Furht B. Deep learning applications for COVID-19. *J Big Data*. 2021;8(1):18-54. https://doi.org/10.1186/s40537-020-00392-9
15. Alakus TB, Turkoglu I. Comparison of deep learning approaches to predict COVID-19 infection. *Chaos Solitons Fractals*. 2020;140:110120. https://doi.org/10.1016/j.chaos.2020.110120
16. Devaraj J, Madurai Elavarasan R, Pugazhendhi R, et al. Forecasting of COVID-19 cases using deep learning models: is it reliable and practically significant?. *Results Phys*. 2021;21:103817. https://doi.org/10.1016/j.rinp.2021.103817
17. Gao J, Sharma R, Qian C, et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *J Am Med Inform Assoc*. 2021;28(4):733-743. https://doi.org/10.1093/jamia/ocaa322
18. Price BS, Khodaverdi M, Halasz A, et al. Predicting increases in COVID-19 incidence to identify locations for targeted testing in West Virginia: a machine learning enhanced approach. *PLoS One*. 2021;16(11):e0259538. https://doi.org/10.1371/journal.pone.0259538
19. Nikparvar B, Rahman M, Hatami F, et al. Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network. *Sci Rep*. 2021;11(1):21715. https://doi.org/10.1038/s41598-021-01119-3
20. Chimmula VK, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals*. 2020;135:109864. https://doi.org/10.1007/s00779-020-01494-0
21. Ma R, Zheng X, Wang P, et al. The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method. *Sci Rep*. 2021;11(1):17421-17424. https://doi.org/10.1038/s41598-021-97037-5
22. Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals*. 2020;140:110212. https://doi.org/10.1016/j.chaos.2020.110212
23. Ghassemi M, Mohamed S. Machine learning and health need better values. *NPJ Digit Med*. 2022;5(1):51-54. https://doi.org/10.1038/s41746-022-00595-9
24. Syrowatka A, Kuznetsova M, Alsubai A, et al. Leveraging artificial intelligence for pandemic preparedness and response: a scoping review to identify key use cases. *NPJ Digit Med*. 2021;4(1):96. https://doi.org/10.1038/s41746-021-00459-8
25. Andeweg SP, Gier B, Eggink D, et al. Protection of COVID-19 vaccination and previous infection against omicron BA.1, BA.2 and Delta SARS-CoV-2 infections. *Nat Commun*. 2022;13(1):4738. https://doi.org/10.1038/s41467-022-31838-8
26. Pisano MB, Sicilia P, Zeballos M, et al. SARS-CoV-2 genomic surveillance enables the identification of Delta/omicron co-infections in Argentina. *Front Virol*. 2022;2. https://doi.org/10.3389/fviro.2022.910839
27. Smoot K, Yang J, Tacker DH, et al. Persistence and protective potential of SARS-CoV-2 antibody levels after COVID-19 vaccination in a West Virginia nursing home cohort. *JAMA Netw*

Open. 2022;5(9):e2231334. https://doi.org/10.1001/jamanetworkopen.2022.31334

28. *Reproduction Numbers from Epidemic Curves*. R package version 2.2-3. 2021. Accessed February 19, 2024. https://CRAN.R-project.org/package=EpiEstim
29. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput*. 2000;12(10):2451-2471. https://doi.org/10.1162/089976600300015015
30. Smagulova K, James AP. A survey on LSTM memristive neural network architectures and applications. *Eur Phys J Spec Top*. 2019;228(10):2313-2324. https://doi.org/10.1140/epjst/e2019-900046-x
31. Hewamalage H, Bergmeir C, Bandara K. Recurrent neural networks for time series forecasting: Current status and future directions. *Int J Forecast*. 2021;37(1):388-427. https://doi.org/10.1016/j.ijforecast.2020.06.008
32. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int J Unc Fuzz Knowl Based Syst*. 1998;06(02):107-116. https://doi.org/10.1142/S0218488598000094
33. Sagheer A, Kotb M. Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. *Sci Rep*. 2019;9(1):19038.
34. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929-1958.
35. Baldi P, Sadowski PJ. Understanding dropout. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2013;26:2814-2822.
36. Garbin C, Zhu X, Marques O. Dropout vs batch normalization: an empirical study of their impact to deep learning. *Multimed Tools Appl*. 2020;79(19-20):12777-12815. https://doi.org/10.1007/s11042-019-08453-9
37. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. OTexts; 2018.
38. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2017;30:4768-4777.
39. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760. https://doi.org/10.1038/s41551-018-0304-0
40. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. https://doi.org/10.1038/s42256-019-0138-9
41. Baptista ML, Goebel KAI, Henriques EM. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artif Intell*. 2022;306:103667. https://doi.org/10.1016/j.artint.2022.103667
42. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*. 2014;41(3):647-665. https://doi.org/10.1007/s10115-013-0679-x
43. Hoi SC, Sahoo D, Lu J, Zhao P. Online learning: a comprehensive survey. *Neurocomputing*. 2021;459(459):249-289. https://doi.org/10.1016/j.neucom.2021.04.112