Fine-Grained Forecasting of COVID-19 Trends at the County Level in the United States

Tzu-Hsi Song[a], Leonardo Clemente[b], Xiang Pan[a], Junbong Jang[a], Mauricio Santillana[b,*], and Kwonmoo Lee[a,*]

[a]Vascular Biology Program and Department of Surgery, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
[b]Department of Physics and Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA

Abstract

The coronavirus (COVID-19) pandemic has profoundly impacted various aspects of daily life, society, healthcare systems, and global health policies. This pandemic has resulted in more than one hundred million people being infected and, unfortunately, the loss of life for many individuals. Although treatment for the coronavirus is now available, effective forecasting of COVID-19 infection is the most importance to aid public health officials in making critical decisions. However, forecasting COVID-19 trends through time-series analysis poses significant challenges due to the data's inherently dynamic, transient, and noise-prone nature. In this study, we have developed the Fine-Grained Infection Forecast Network (FIGI-Net) model, which provides accurate forecasts of COVID-19 trends up to two weeks in advance. FIGI-Net addresses the current limitations in COVID-19 forecasting by leveraging fine-grained county-level data and a stacked bidirectional LSTM structure. We employ a pre-trained model to capture essential global infection patterns. Subsequently, these pre-trained parameters were transferred to train localized sub-models for county clusters exhibiting comparable infection dynamics. This model adeptly handles sudden changes and rapid fluctuations in data, frequently observed across various times and locations of county-level data, ultimately improving the accuracy of COVID-19 infection forecasting at the county, state, and national levels. FIGI-Net model demonstrated significant improvement over other deep learning-based models and state-of-the-art COVID-19 forecasting models, evident in various standard evaluation metrics. Notably, FIGI-Net model excels at forecasting the direction of infection trends, especially during the initial phases of different COVID-19 outbreak waves. Our study underscores the effectiveness and superiority of our time-series deep learning-based methods in addressing dynamic and sudden changes in infection numbers over short-term time periods. These capabilities facilitate efficient public health management and the early implementation of COVID-19 transmission prevention measures.

1 Introduction

Since 2020, the Coronavirus disease 2019 (COVID-19) has rapidly spread from several countries to the global stage, prompting the World Health Organization (WHO) to declare it a global pandemic [12]. With over half a billion infections and more than 6 million deaths recorded worldwide [14], COVID-19 has significantly shaped society, national healthcare systems, and the global economy. Due to its fast mutation rate, new outbreaks of COVID-19 continue to emerge within a span of a few months, making forecasting the trends of these COVID-19 waves a crucial task that helps public health officials to regulate public healthcare policies and efficiently manage medical resources to prevent and control future outbreaks [16].

*Corresponding authors: Kwonmoo.Lee@childrens.harvard.edu (Kwonmoo Lee), m.santillana@northeastern.edu (Mauricio Santillana)

Epidemic forecasting presents a formidable challenge, driving the development of numerous methods to address this complexity. With the advent of the COVID-19 pandemic, the need for efficient preventive measures and accurate forecasting models tailored to this emerging infectious disease became increasingly critical. Within the field of epidemiology, Susceptible–Infected–Removed (SIR) models and their variations are extensively utilized to assess the spread of infections [25, 22]. These epidemiological models allow us to deduce crucial parameters such as infection and death rates, enabling the forecast of disease transmission trends. Several COVID-19 studies have demonstrated practical insights and forecasts by employing SIR-like models in various regions worldwide [28, 26, 29, 13, 33, 2, 35, 4, 20, 9, 39, 44, 10]. However, their ensemble averaging approaches often fall short in capturing essential local heterogeneities in infection dynamics. In contrast, data-driven methods may excel in this context by incorporating the influence of diverse local policies, new variants, demography, and socioeconomic factors, which are challenging to quantify using SIR-like models. Notably, the United States displays highly diverse vaccination rates, adding complexity to the accurate characterization and prediction of disease spread. Hence, alternative data-driven models could offer complementary insights into disease dynamics, enhancing our forecasting of the ongoing pandemic [43].

Hence, machine learning models have been extensively applied in disease forecasting [8, 40, 41]. For instance, Sujath *et al.* conducted a comparison of linear regression, multilayer perceptron (MLP), and vector autoregression, with the MLP model demonstrating superior performance [42]. Ardabili *et al.* integrated genetic algorithms with SIR and SEIR models to predict outbreak trends [3]. Hernandez *et al.* utilized the Auto-regressive Integrated Moving Average (ARIMA) [19] model with polynomial functions for global COVID infection trend predictions, using cumulative datasets [21]. Additionally, Lu *et al.* compared and integrated various approaches, combining statistical linear and nonlinear models to estimate the cumulative number of weekly confirmed cases [31]. Many of these models rely on assumptions about the constancy of raw data or adherence to specific epidemic behaviors. However, the dynamics of COVID-19 are in constant flux, influenced by factors such as virus characteristics and intervention policies. Consequently, COVID-19 forecasting has become an exceptionally challenging endeavor.

Deep learning's feature-learning ability holds promise for addressing challenges stemming from the local heterogeneity of infection dynamics. Several studies have utilized deep learning methods to predict COVID-19 data, incorporating related time series models such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and hybrid approaches that combine deep learning with traditional methods. Rodriguez *et al.* utilized a feed-forward network to generate short-term forecasts of COVID-19, demonstrating the responsive capacity of Network models to adapt to sparse data situations. Bandyopadhyay *et al.* used LSTM with a gate circulation network to estimate COVID-19 cases [15], while Huang *et al.* employed CNNs to predict cumulative COVID-19 deaths [27]. Ruifang *et al.* combined LSTM with a Markov method for national cumulative COVID-19 predictions [32]. ArunKumar *et al.* compared the performance of deep learning models (GRU and LSTM) with traditional models like ARIMA and SARIMA, concluding that deep learning models were better suited for non-linear datasets [5]. Additionally, Transformer models, another category of deep learning methods, have been utilized in COVID-19 forecasting. Soumyanil *et al.* developed a graph transformer network with synchronous temporal and spatial information [6], while Kapoor *et al.* used a similar approach to predict COVID-19 confirmed cases [24]. However, these models were based on limited country or state-level datasets, where the local heterogeneity of COVID-19 infections was averaged out. Consequently, despite the impressive capabilities of deep learning, their forecasting results were restricted, especially during the rapidly changing outbreak of the COVID-19 variants such as Omicron.

In response to these challenges, we present a novel deep learning approach that harnesses extended time series data covering the entire span of the COVID-19 pandemic in the United States, including the Omicron wave. Our model, built on the foundation of Long Short-Term Memory (LSTM) networks and incorporating stacked bidirectional components, capitalizes on temporal relation aware to adeptly manage abrupt shifts in infection dynamics. This ensures precise short-term predictions vital for shaping future epidemic prevention and control strategies. Our method, named FIGI-Net (Fine-Grained Infection Forecast Network), delves into fine-grained time series COVID-19 infection data at the county level in the U.S. Initially, we segmented historical daily confirmed cases and deaths into distinct half-month periods, allowing us to train the global biLSTM model using infection case data of all U.S. counties. This approach leverages LSTM capabilities with a stacked bidirectional component,

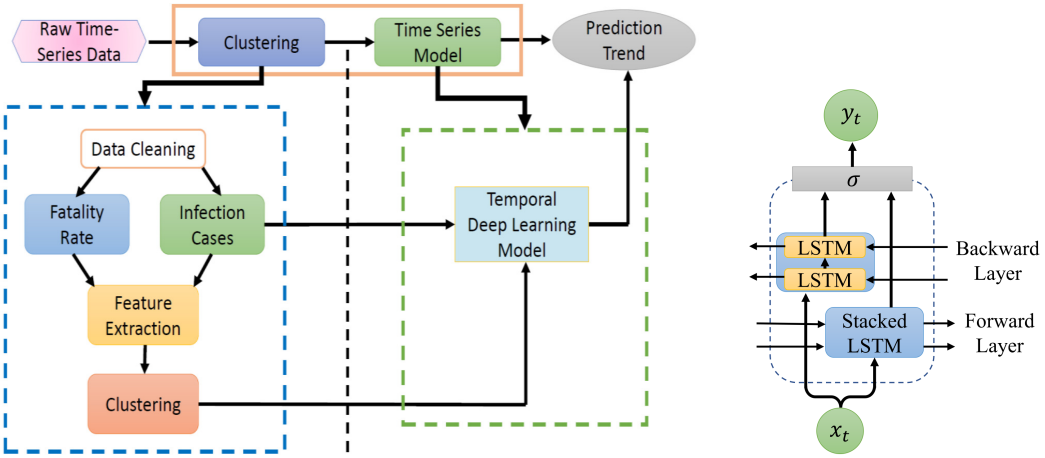

Figure 1: Our proposed model, FIGI-Net, architecture for COVID-19 infection prediction. (a) A diagram that visualizes the our methodology's framework, consisting of a clustering and a time series deep learning component. In the clustering section, auto-correlation and cross-correlation were used to extract similar features from the pre-cleaning infection cases and fatality rate data. Then, we applied a clustering technique to the extracted features as a means to identify similar COVID-19 dynamics. The temporal deep learning model utilized the given clusters along with the infection case data to predict the trend of COVID-19. (b) The architecture of the proposed bidirectional stacked LSTM model. This model learns the feature dependencies of input sequential data in both forward and backward directions and can effectively deal with short-term state changes. The σ function represents the merging function.

enhancing the model's capacity to discern diverse global infection trends. Subsequently, upon identifying county clusters with similar COVID-19 temporal patterns, we applied transfer learning from the global biLSTM model to train specific cluster-based biLSTM models. This cluster transfer learning concept preserves the model's inherent capabilities while empowering it to swiftly adapt to short-term trend changes, refining forecasting accuracy. Using our proposed FIGI-Net, we conducted predictions for future COVID-19 infection cases, encompassing both short-term and long-term forecasts spanning from the next day to the next 15 days. The key contributions of our framework are summarized as follows:

- We introduce FIGI-Net (Fine-Grained Infection Forecast Network), a deep learning pipeline that leverages COVID-19 infection time series data of U.S. counties, and is capable of forecasting COVID-19 confirmed cases up to 15 days ahead. Harnessing bidirectional temporal feature learning and transfer learning techniques, our model was trained with geographical clusters of COVID-19 temporal data. The structure of the proposed framework is detailed in Figure 1 and more details are presented in section 1.
- Focusing on US county-level COVID-19 time series data, our study demonstrates the FIGI-Net's remarkable efficiency and accuracy in handling rapid fluctuations in COVID-19 activity, both in short-term (up to 1 day) and long-term (up to 2 weeks) forecasting scenarios.
- Furthermore, our methodology was extended to national and state-level forecasts using weekly time periods. The results highlight the superiority of our FIGI-Net, showcasing more than a 30% reduction in errors during critical time periods compared to other state-of-the-art models designed for COVID-19 infection forecasting tasks.

Results

We implemented a collection of machine-learning based models to generate out-of-sample predictions for the number of COVID-19 confirmed cases, as reported by the Centers for Disease Control and

Prevention (CDC), for the time period between April 15$^{\text{th}}$, 2020 and April 15$^{\text{th}}$, 2022. These models included the model we propose: FIGI-Net, as well as Autoregressive statistical models, a collection of neural network based models (GRU, LSTM) and a "naive" (persistence model) to be used as a baseline.

Our models retrospectively produced daily forecasts for multiple time horizons h, ranging from 1 to 15 days ahead ($h = 1, 2...15$). Visual representations of our predictions, alongside the actual observed COVID-19 confirmed cases, are presented in Figure 2 and 3 for county, state, and national level. We evaluated our model's performance in each time horizon h using the root mean square error (RMSE), relative RMSE (RRMSE), and mean absolute percentage error (MAPE) metrics, as detailed in Section 1. We compared the prediction performance of our models with all other models we implemented as well as with state-of-the-art forecasting methodologies reported by the CDC (a comprehensive list of these models is provided in Section 1). Finally, we assessed our model's ability to anticipate the onset of outbreaks during critical time periods (periods where confirmed COVID-19 cases were observed to grow rapidly).

FIGI-Net Forecasting Performance at County Level

For each forecasting horizon h, we computed FIGI-Net's RMSE and RRMSE scores across all counties, over the whole time period. Table 1 shows the metric scores (and percentage of error reduction with respect to the persistence model) of all the models for 1-day, 7-day and 14-day horizons. Our results demonstrate that FIGI-Net model has the greatest error reduction across all horizons (87%, 50%, and 42% RMSE reduction, correspondingly), Followed by the biLSTM model (83% in horizon 1-day, 49% in horizon 7-day, and 41% in horizon 14-day RMSE reduction). To assess the significance of these error reductions, we performed the two-sided Wilcoxon rank sum test [37] over the entire outcomes of tasks.

In Figure 2.A, we visualize the forecasting ability of each model for both the 1-day and 7-day ahead task, including classic models such as persistence (a naive rule stating $y_{t+1} = y_t$) and Autoregressive models(AR) [7], and deep learning-based models such as GRU or LSTM architectures. We used the median RMSE score for each model as a means to order them in decreasing order (leftmost model with the worst performance, and rightmost with the best). Our first observation is that FIGI-Net scored the lowest median RMSE and relative RMSE scores across all models for both the 1-day ahead and the 7-day ahead prediction task (approximately 13.97% at 1-day ahead, and 36.11% at 7 day ahead), followed by deep learning models with bidirectional components (TC-biLSTM with a reduction of 22.39% and 46.13% and bi-SLSTM with a reduction of 18.77% and 37.6%, respectively). Generally speaking, all network based models improved over Persistence (with an error of 96.51% and 70.05%) and the Autoreggressive models (94.67% and 121.35%). For a detailed description of the performance of each model, please refer to Table 1.

Figure 2.B focuses on the performance of FIGI-Net against Persistence. For each time horizon, we generated a violin plot to visualize the RMSE scores of FIGI-Net (in blue) and Persistence (in orange) across all counties. Our results show that FIGI-Net has a higher concentration of scores between the 0-20 range across all time-horizons, in comparison to Persistence, where the scores of the orange distribution are widely spread. The main error difference between FIGI-Net and Persistence occur at the horizon 1, with a mean RMSE score of 5.472 , in comparison to 36.323 from Persistence (a 84.93% reduction).

Figure 2.C showcases a visualization of the forecasts of our model for the county of Suffolk, Massachusetts, for the 1-day, 7-day and 14-day ahead tasks. Additionally to the full time period of the experiment, 3 periods from May to August 2020, October 2020 to February 2021, and December 2021 to March 2022, are also displayed. FIGI-Net accurately forecasted the daily infection trends in 1-day and 7-day ahead horizons in diverse counties, even when these counties exhibited contrasting infection trends, as shown for the country of Suffolk during the initial outbreak stages. However, we observed that the 14-day ahead forecasting trend yielded larger errors, particularly in cases where the infection numbers fluctuated significantly.

Forecasting Performance at State Level

Similar to our county level experiment, we compared the performance of FIGI-Net against state-of-the-art models for the state-level geographical resolution. For FIGI-Net, which is a model that leverages

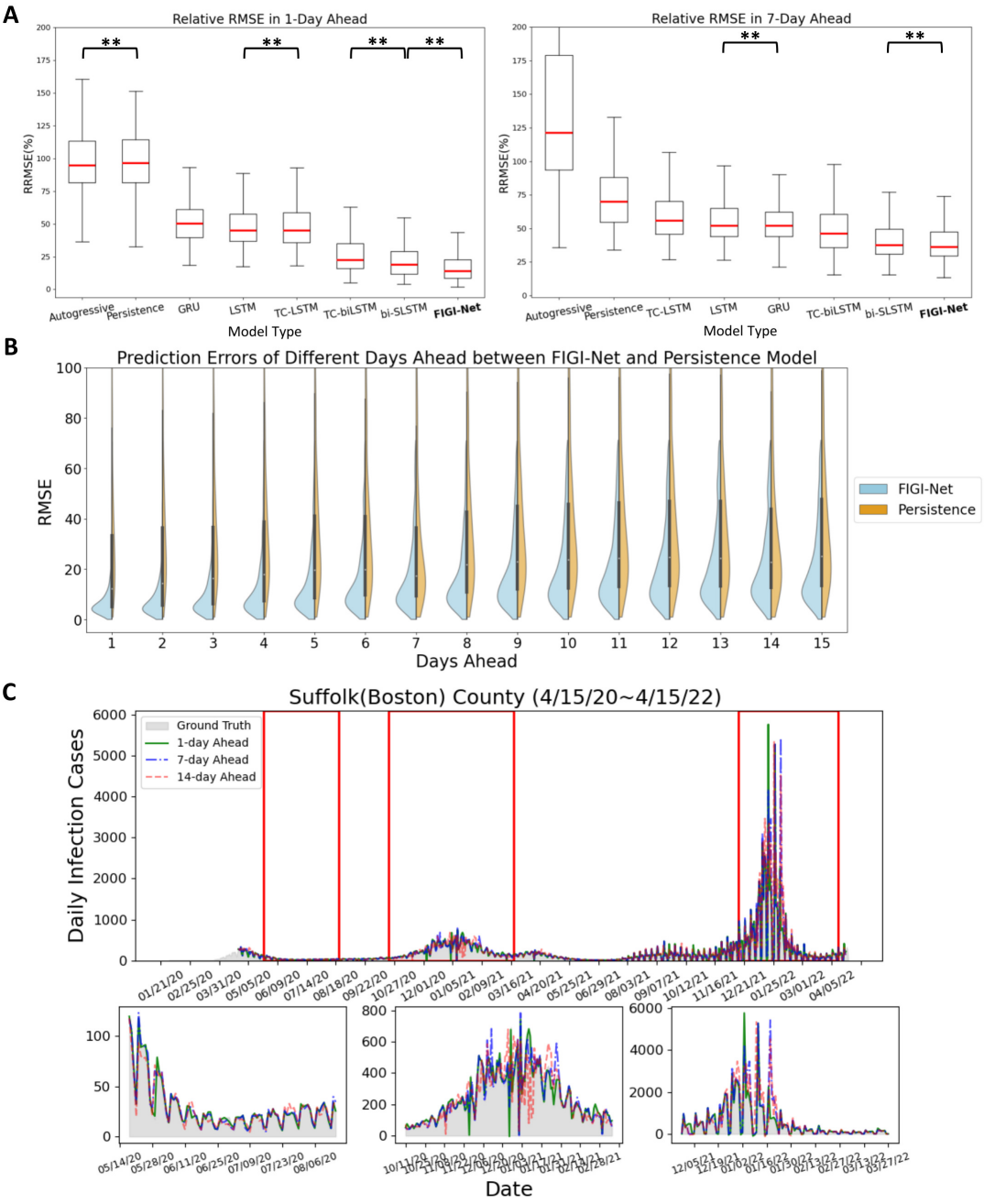

Figure 2: A summary of the performance of FIGI-Net at the county level, for the 1-day and 7-day ahead horizon tasks. (A) Performance of each model over all 3143 counties presented as a boxplot. The median is highlighted in red along with the 5^{th} and 95^{th} percentile whiskers. The models are ordered in decreasing order, with the most accurate model (lowest RRMSE) appearing on the rightmost side. Notably, FIGI-Net exhibits the lowest RRMSE compared to other models for both tasks, as confirmed by the two-sided Wilcoxon rank sum test. (B) Performance comparison of FIGI-Net against Persistence (our baseline model).Comparison is shown as a set of violin plots across the different time horizons. Our model consistently displays a narrower distribution of prediction errors (RMSE) compared to the Persistence model. (C) shows Daily prediction results of our model in 1-day, 7-day and 14-day horizons in Suffolk county. The example demonstrates our model's ability to provide highly accurate predictions for diverse locations and various time periods.
**: p-value ¡ 0.001

Table 1: Performance Metrics with Error Reduction for the 1-day, 7-day and 14-day ahead tasks at County level.

Model	RMSE						RRMSE(%)					
	1 day		7^{th} day		14^{th} day		1 day		7^{th} day		14^{th} day	
	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction
Persistence	79±148.53	—	60.25±109.51	—	84.53±150.93	—	96.31±26.65	—	72.97±22.37	—	102.49±19.95	—
Autoregressive	88.25±182.47	-11.71%	$65700\pm1.61x10^5$	< -100%	$4.08x10^8\pm1.21x10^{10}$	< -100%	104.37±73.16	-8.37%	$65824\pm1.37x10^5$	< -100%	$3.61x10^8\pm1.02x10^{10}$	< -100%
GRU	42.19±83.86	46.59%	44.98±87.29	25.34%	52.45±97.63	37.95%	52.04±17.26	45.97%	56.56±23.46	22.49%	65.78±27.41	35.82%
LSTM	38.76±75.7	50.94%	44.5±85.34	26.14%	52.17±96.96	38.28%	48.9±17.91	49.23%	56.1±17	23.12%	65.34±26.65	36.25%
TC-LSTM	40.79±88.69	48.37%	47.31±91.73	21.48%	54.03±100.74	36.08%	48.72±18.19	49.41%	60.11±21.26	17.62%	68.19±25.31	33.47%
TC-biLSTM	21.17±49.44	73.2%	37.23±63.44	38.2%	55.89±92.67	33.88%	29.56±20.12	69.3%	49.84±18.77	31.7%	71.34±24.8	30.39%
biLSTM	13.43±25.79	83%	30.53±47.47	49.33%	49.2±77.94	41.8%	25.29±20.1	73.74%	42.26±16.28	42.09%	63.39±21.26	38.15%
FIGI-Net	**9.97±16.85**	**87.38%**	**29.81±45.11**	**50.52%**	**48.72±79.72**	**42.36%**	**20.58±19.21**	**78.63%**	**40.74±16.25**	**44.17%**	**62.16±20.26**	**39.35%**

Table 2: Performance Metrics with Error Reduction for the 1-day, 7-day and 14-day ahead tasks at State Level.

Model	RMSE						RRMSE(%)					
	1 day		7^{th} day		14^{th} day		1 day		7^{th} day		14^{th} day	
	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction
Persistence	4236.4±6492.03	—	3201.06±4571.6	—	3896±5151.15	—	88.77±29.51	—	66.47±21.17	—	81.9±15.5	—
Autoregressive	4527.78±6824.48	-6.87%	$3.972x10^6\pm1.873x10^7$	< -100%	$2.51x10^{10}\pm1.21x10^{11}$	< -100%	94.02±39.97	-5.92%	59757.5±2.85x10^5	< -100%	$3.96x10^8\pm2.09x10^9$	< -100%
GRU	2079.54±3193.51	50.91%	2179.06±3285.63	31.93%	2634.01±3782.2	32.4%	45.2±16.17	49.09%	46.79±14.9	29.6%	55±13.97	32.84%
LSTM	1856.7±2801.76	56.17%	2161.51±3220.82	32.48%	2631.33±3753.86	32.46%	41.06±15.94	53.74%	46.24±14.68	30.43%	55.01±15.3	32.84%
TC-LSTM	1922.61±3417.42	54.62%	2244.81±3524.86	29.87%	2636.06±3949.36	32.34%	40.49±16.53	54.39%	47.97±14.32	27.83%	54.64±13.62	33.28%
TC-biLSTM	864.21±1289.96	79.6%	1666.07±1943.54	47.95%	2615.52±3488.19	32.87%	20.77±16.16	76.6%	37.11±13.9	44.16%	55.59±15.65	32.13%
biLSTM	606.73±858.48	85.68%	1411.99±1513.6	55.89%	2318.1±2706.3	40.5%	15.88±14.55	82.11%	32.6±13.41	50.96%	51.82±16.89	36.73%
FIGI-Net	**509±741.08**	**87.99%**	**1408.26±1544.87**	**56.01%**	**2292.83±2924.04**	**41.15%**	**12.56±13.06**	**85.85%**	**31.68±13.57**	**52.33%**	**50.33±16.31**	**38.55%**

Table 3: Performance Metrics with Error Reduction for the 1-day, 7-day and 14-day ahead tasks at Nation Level.

Model	RMSE						RRMSE(%)					
	1 day		7^{th} day		14^{th} day		1 day		7^{th} day		14^{th} day	
	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction	Error	Reduction
Persistence	94151.72	—	71901.03	—	109661.49	—	51.93	—	39.65	—	60.48	—
Autoregressive	91127.52	3.21%	$1.334x10^8$	< -100%	$8.8278x10^{11}$	< -100%	50.26	3.22%	73574.39	< -100%	$4.869x10^8$	< -100%
GRU	41298.1	56.14%	38823.91	46%	57823.73	47.27%	22.78	56.13%	21.41	46%	31.89	47.27%
LSTM	32536.62	65.44%	39022.06	45.73%	60314.95	45%	17.94	65.45%	21.52	45.73%	33.26	45.01%
TC-LSTM	37139.2	60.55%	41291.44	42.57%	55759.71	49.15%	20.48	60.56%	22.77	42.57%	30.75	49.16%
TC-biLSTM	16426.44	82.55%	35465.95	50.67%	62286.61	43.2%	9.06	82.55%	19.56	50.67%	34.35	43.2%
biLSTM	13569.52	85.59%	**27533.97**	**61.71%**	51447.97	53.08%	7.48	85.6%	**15.18**	**61.72%**	28.37	53.09%
FIGI-Net	**12726.38**	**86.48%**	28045.28	60.99%	**50654.95**	**53.8%**	**7.02**	**86.48%**	15.47	60.98%	**27.94**	**53.8%**

high volumes county-level activity as part of its design, we decided to aggregate the county level forecasts into state level. For the other models, we used the same approach, as FIGI-Net, to obtain state-level outcomes. Then, we computed the RRMSE for each state across 1-day, 7-day, and 14-day ahead horizons, as shown in Figure 3.A. Our results, also summarized in table 2, show that FIGI-Net was able to score 1408.25±1544.87 and 2292.83±2924.04 in terms of RMSE for the 7-day and 14-day horizon, correspondingly. On the other hand, the score of the persistence model was 3201.06±4571.6 and 3896±5151.15 (resulting in an error reduction of 56.01%, and 41.15% in each case). Next to FIGI-Net, we can see the biLSTM (with a score of 1411.99±1513.6 and 2318.1±2706.3) and TC-biLSTM models (1666.07±1943.54 and 2615.52±3488.19 in RMSE).

During the early stages of the pandemic, we observe the RRMSE values of FIGI-Net were notably higher. For instance, based on Figure 3.A, Missouri, Montana, and Nevada displayed larger prediction errors between March to May 2021, as shown from the deep red color in the center of the RRMSE matrix. During this period, these states experienced a significant spike in COVID-19 activity, deviating from both previous months and the overall national trend. Figure 3.B represents country-level COVID-19 official reports contrasted with our model predictions across three different horizons. Similar to the observations in Figure 2.C, our predictions were highly accurate at the 1-day and 7-day ahead horizons, with the RRMSE of 7.02% and 15.47%, respectively, compared to the national official reports. However, at the 14-day horizon, the discrepancies between predicted values and ground truth grew during high infection periods, resulting in an RRMSE of 27.94%.

National Level Performance

Additionally, we analyzed the daily performance of FIGI-Net at the national level across three main outbreak waves: the Alpha-variant wave (October 15^{th}, 2020 to March 15^{th}, 2021), Delta-variant wave (July 15^{th}, 2021 to November 1^{st}, 2021), and Omicron-variant outbreak wave (November 15^{th}, 2021 to March 31^{st}, 2022). By aggregating our county-level forecasts, we depicted the national prediction trends during these waves (Figure 3.C). Our model consistently demonstrated an accurate prediction direction, maintaining an average confidence interval of 8.05 infection cases during the Alpha-variant

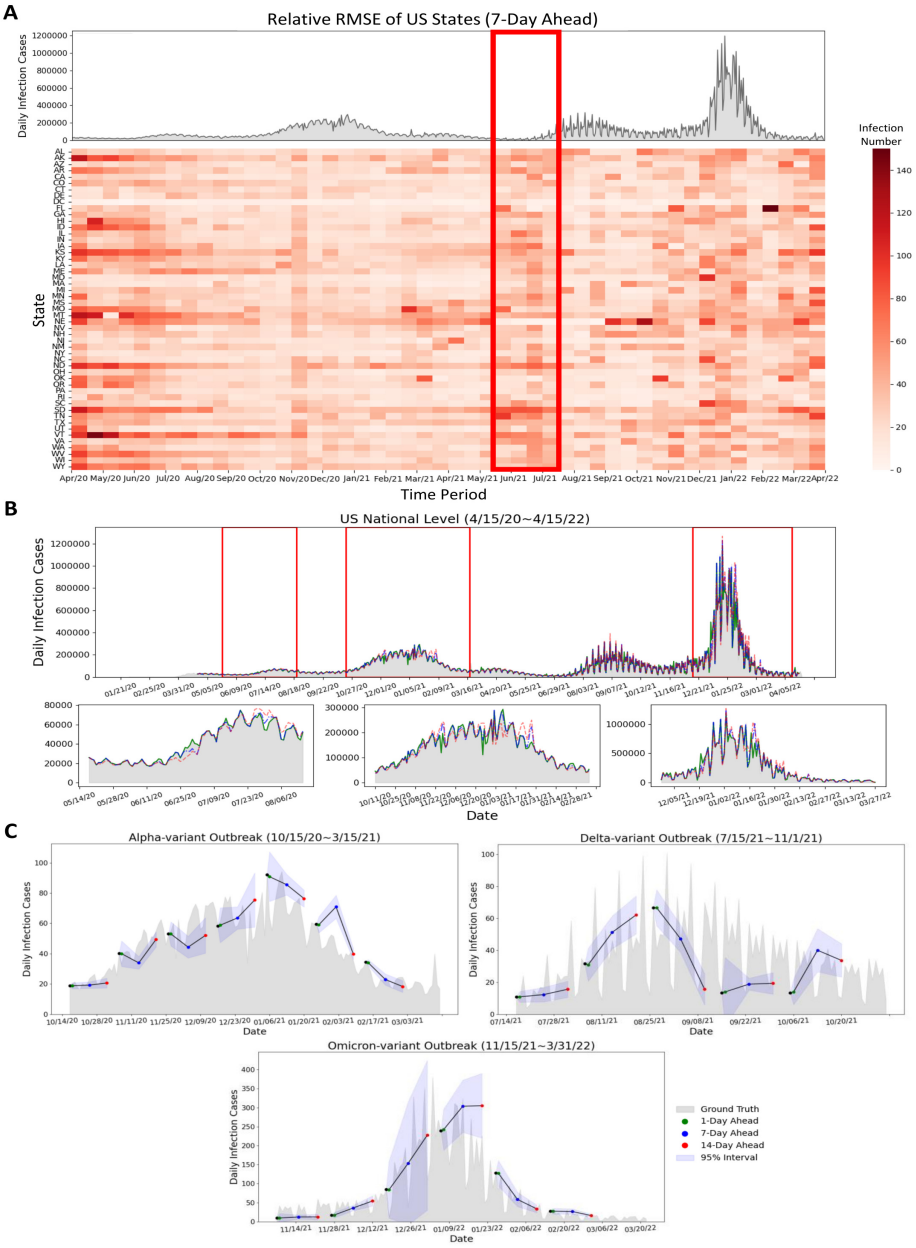

Figure 3: Summary of forecasting results at the state and national level. (A) Relative RMSE performance among US states in the 7-day horizon at the state level is determined by the average relative RMSE of the last 7 days of each time period, compared to the national reported infection trend. The relative RMSE increased during the early period (April 2020 to May 2020) of the first upward trend, the time period (June 2021 to July 2021) before the Delta outbreak, and the increasing period (December 2021 to January 2022) of the Omicron COVID-19 outbreak. Missouri, Montana, and Nebraska have large relative RMSE values during March 2021 to May 2021. We can observe that the RRMSE errors often increase before the early stage of the next outbreaks or when the infection trend rapidly increases (red rectangle) (B) Daily prediction infection trends during different days ahead at the national level. It can be observed that the daily predicted infection trends at 1-day and 7-day ahead horizons show similarity to the reported data, while the 14-day ahead trend exhibits some fluctuations. (C) Daily prediction trends of the proposed model during Alpha-variant, Delta-variant, and Omicron-variant outbreaks with different days ahead predictions. The proposed model can provide a curve of predicted trends matching the observed report. The interval range became larger when the trend of Omicron-variant outbreak increased.

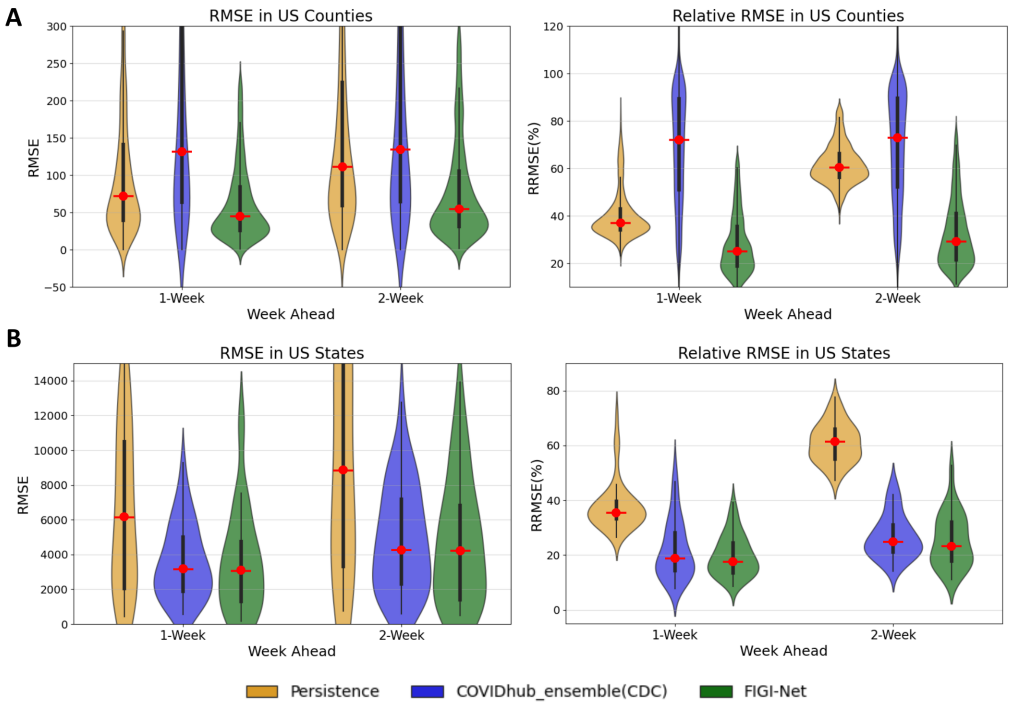

Figure 4: Comparison of prediction performance among persistence model, CDC model, and our proposed model. (A) The RMSE and relative RMSE values of the three models at the county level. (B) The comparison of prediction errors among the three models at the state level. Our proposed model outperforms the persistence model and CDC model in terms of lower prediction RRMSE errors. This indicates its enhanced capability to capture the complex dynamics of COVID-19 infection spread, with approximate 10.7% averaged reduction in errors observed across various prediction horizons.

Table 4: Comparison of FIGI-Net with CDC and Persistence Models at County Level

Model	1 week				2 week			
	RMSE		RRMSE(%)		RMSE		RRMSE(%)	
	Value	Error reduction	Value	Error reduction	Value	Error reduction	Value	Error reduction
Persistence	170.19±288.67	6.67%	42.33±13.92	29.96%	274±488.52	51.99%	64±11.38	42.22%
CDC_ensemble	449.88±1252.27	77.76%	91.65±77.19	67.65%	423.56±1058.05	68.94%	87.52±65.32	57.75%
FIGI-Net	**100.04±146.29**	—	**29.65±19.55**	—	**131.54±216.38**	—	**36.98±26.47**	—

Table 5: Comparison of FIGI-Net with CDC and Persistence Models at State Level

Model	1-week				2-week			
	RMSE		RRMSE(%)		RMSE		RRMSE(%)	
	Value	Error reduction	Value	Error reduction	Value	Error reduction	Value	Error reduction
Persistence	9405.24±11099.01	49.37%	39.13±10.67	45.62%	15746.93±19521.28	61.99%	62.95±9.49	55.85%
CDC_ensemble	**4627.1±4629.7**	—	25.68±21.92	17.13%	7085.08±8549.48	15.52%	31.93±18.45	12.97%
FIGI-Net	4761.8±4744.92	2.83%	**21.28±10.58**	—	**5985.21±6183.14**	—	**27.79±14.21**	—

and Delta-variant waves. During the Omicron-variant wave, the forecasting trend exhibited a wide range of confidence intervals. However, our model still forecasted consistent trends up to 14-day horizon. Also, Table 3 shows the performance between FIGI-Net and other models at nation level. Comparing to persistence model and other models, FIGI-Net has better forecasts and can lead to a 86.48% reduction at 1 day horizon, a 60.98% reduction at 7 day horizon, and a 53.8% reduction at 14 day horizon in RRMSE score. These above outcomes underscore the robustness of our FIGI-Net model in addressing substantial variations in infection numbers.

Comparison between FIGI-Net and CDC Ensemble Model in COVID-19 Prediction

As an additional benchmark, we compared the performance of our proposed FIGI-Net model with the COVIDhub_ensemble model [11] (also known as the CDC model). The CDC model employs an ensemble methodology that combines the output of several disease surveillance teams across the United States, generating forecasts for the number of COVID-19 infections at the county, state and national levels. We collected the 1-week and 2-week ahead forecasts of the CDC model at county and state level, and compared them against FIGI-Net's forecasts. Given the CDC model is an aggregated forecast (i.e. the total number of reported activity over the next 7 and 14 days, rather than a daily forecast over the same periods), we aggregated our daily predictions for the 1 to 7-day and 1 to 14-day horizons to facilitate a fair comparison between both models. A persistence model of this task was also included as baseline.

As depicted in Figure 4.A, the CDC model exhibited significantly higher average RMSE and RRMSE values at the county level compared to our FIGI-Net model. Our model achieved an approximate reduction of 51% in averaged RMSE and 49.4% in averaged RRMSE over the 1 and 2-week (7 and 14-day) ahead horizons, respectively (see Table 4). At the state level (Figure 4.B), our proposed model consistently maintained the averaged reduction of 31.01% RMSE value and 32.89% RRMSE value (compared to the CDC ensemble and persistence models, as shown in Table 5). Notably, the CDC model demonstrated better performance than the persistence model in terms of lower error predictions for both 1 and 2-week horizons.

Additional to the state-of-the-art comparison analysis previously shown, we performed a comparative analysis between the COVID-19 forecasts of FIGI-Net with other forecasting models officially reported on the CDC website [11]. We selected models, including Microsoft-Deep, JHU_CSSE_DECOM, Karlen-pypm, CovidAnalytics-DELPHI, and MIT_ISOLAT-Mixtures, that provided sufficient infection prediction outcomes for 1-week and 2-week ahead horizons [11]. Following the CDC weekly reporting criteria [1], we aggregated the daily prediction cases into weekly prediction values. Specifically, we focused on three reported infection wave time periods and presented 1-week and 2-week prediction results of our model at the national level alongside those of other models (Figure 6.A). Our findings revealed that most models can accurately predict infection numbers during decreasing trend, but struggle to forecast the correct trend direction and infection numbers during increasing trend. Interestingly, our FIGI-Net model correctly predicted the increasing direction of the infection trend before the wave began, from November 2021 to March 2023. Furthermore, Figure 6.B illustrates the examples of the predicted infection trends in Massachusetts and New York states by our FIGI-Net model and other forecasting models in 1-week and 2-week horizons. The results showed that the our model's infection prediction trends are much closer to the reported data across different weeks ahead at the state level, and most of other models predicted accurate infection numbers when the infection trend increases. It is important to note that some models did not provide outcomes at some weekly time points,leading to 0 values in those models, which were subsequently removed in further comparison and analysis.

Geographical Distribution of COVID-19 Infection Predictions of US Counties

In this section, we conducted a geo-spatial analysis with the objective to identify possible geographical patterns in the performance of FIGI-Net. Figure 5 illustrates different geographical infection distribution maps and relative RMSE errors of US counties for the 7-day ahead task, spanning various time periods. Specifically, we chose three critical time frames: the first halves of August 2020, August 2021, and January 2022, representing periods just before the Alpha-variant, Delta-variant, and Omicron-variant waves began, respectively. For each period, last 7-days' forecasting predictions were averaged to create these geographical infection and error distribution maps. FIGI-Net accurately successfully predicted the COVID-19 activity shown in the 7-day ahead prediction map (Figure 5(b), (e), and (h)) compared to the observation maps (Figure 5(a), (d), and (g)). However, some counties had larger errors, as observed from the relative RMSE maps (Figure 5(c), (f), and (i)). During these time periods, counties in Kansas and Louisiana displayed higher errors despite low or mild infection levels. Moreover, the counties in Oklahoma, Iowa, Michigan, and Florida exhibited larger prediction errors preceding the Delta-variant wave. Particularly in Florida, while the epidemic situation was severe, errors increased. Our model demonstrated higher accuracy and lower RRMSE values in the west coast and northeast

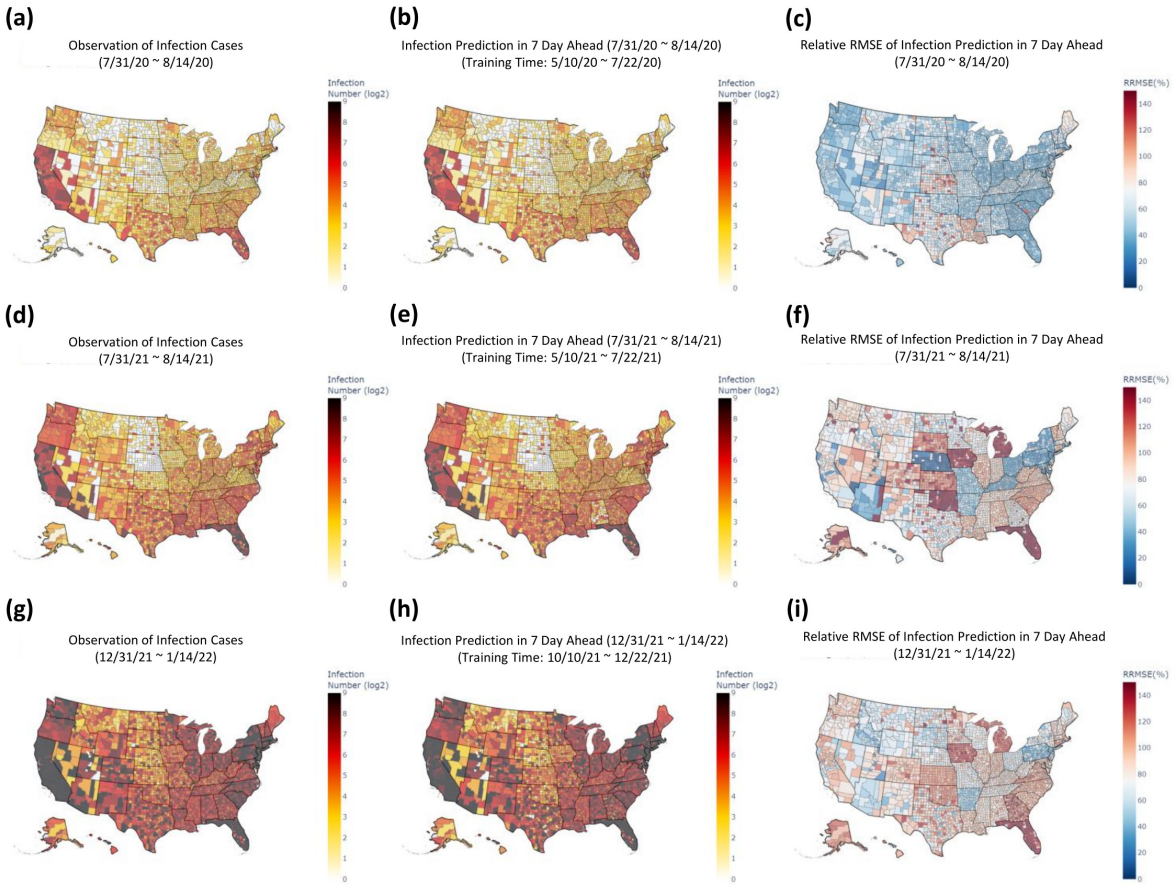

Figure 5: Geographical distribution of infection prediction errors of the proposed model during different time periods. The figure shows the infection prediction distribution of the proposed model during different time periods at county level. The early half of August 2020, August 2021, and January 2022 are represented in (a)-(c), (d)-(f), and (g)-(i), respectively. (a), (d), and (g) show the observed daily confirmed cases. (b), (e), and (h) represent the 7-day (1-week) ahead prediction results, and (c), (f), and (i) are the corresponding RRMSE score. The proposed model provided the predictions that are similar to the observed daily reports of infection cases and had low RRMSE values in most counties. However, it had higher RRMSE values in the counties where the number of infection cases rapidly increased. Additionally, our model indicated higher RRMSE values in the counties of Kansas and Louisiana at these time periods. The counties in Michigan and Florida represented much higher RRMSE values during time periods (f) and (i) when the status of infections in these two states were severe.

regions during these time periods (see, for example, Figure 5(c), (f), (i)), while mid-west and south regions of the U.S. tended to display higher errors as the pandemic progressed (refer to Movie S.1 for the details of geographical distribution prediction and error maps in 1-day, 7-day, and 14-day ahead across all time periods from April 2020 to April 2022).

Comparative Analysis of COVID-19 Forecasting Models During Critical Time Periods

Given identifying the beginning of a major outbreak is a crucial task in digital disease forecasting, We assessed to the performance of our FIGI-Net model in early COVID-19 prevention and forecasting by measuring its ability to anticipate critical time periods marked by exponential growth of COVID-19 infection cases. We compared our model's performance with other state-of-the-art COVID-19 forecasting models during this critical time period, using weekly data from July 20^{th}, 2020 to April 11^{th}, 2022. First, we identified these 'critical' time periods as periods where the trend λ of COVID-19

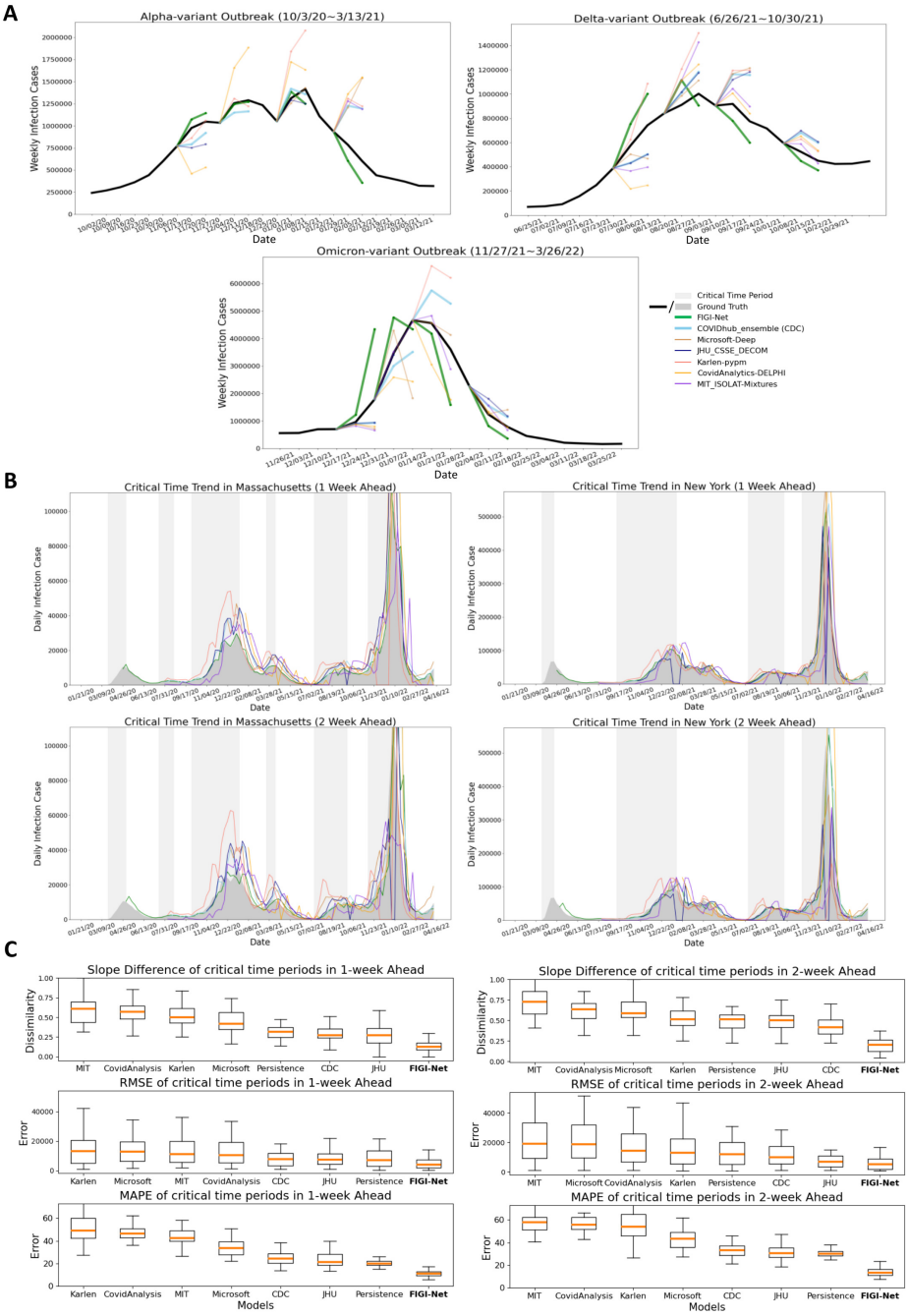

Figure 6: Comparison of FIGI-Net model with other state-of-the-art prediction models in predicting weekly and critical time infections. (A) Comparison of the forecasting results among different forecasting models during three different outbreak time periods at the national level. (B) Examples of COVID infection prediction trends of different state-of-the-art forecasting models at the state level. The critical time period, which indicates a significant increase in COVID infections, is highlighted in light grey color. (C) Performance evaluation of the forecasting methods during the critical time periods of COVID infection in 1-week and 2-week horizons across the states. Slope Similarity, RRMSE, and MAPE were measured to assess the prediction number and trend accuracy of each model. Our proposed FIGI-Net model provided lower prediction errors in both 1-week and 2-week horizons during the critical time and may efficiently forecast the infection number and trend direction before the severe transmission of COVID-19. Here we also ranked them from high to low evaluation or error values according to the median values.

activity (estimated as the coefficient of a linear model $y_t = \lambda y_{t-1}$) remained above 1, indicating a sustained multiplicative growth for an extended period. Examples of such periods for Massachusetts and New York states are shown in Figure 6.B. We compared the ground truth and predicted results of the models using slope similarity, RMSE, and MAPE values (see Section 4.4 for a definition to calculate slope similarity). Our FIGI-Net model efficiently and accurately predict infection case numbers and trend direction for 1-week and 2-week horizons during critical time periods (Figure 6.C). Tables 6 and 7 presented the forecasting performance details among the prediction models and the statistical significance between our model and the others. FIGI-Net model improved the RMSE value by at least 50% reduction in 1-week ahead and at least 22% reduction in RMSE in 2-week ahead forecasts. Additionally, our model achieved at least a 60% MAPE reduction in both week horizons and exhibited 70% similarity in the slope of infection trend. These results indicate that FIGI-Net model can effectively adapt and refine forecasting trends when the pandemic intensities suddenly.

Discussion

In this work, we presented FIGI-Net, a deep learning-based model that utilizes fine-grained county level infection time-series data for short-term forecasting up to two weeks. We conducted an experimental analysis that compared the forecasting ability of FIGI-Net against state-of-the-art methodologies, including autoreggressive models, recurrent network architectures such as GRU and LSTMs, and more advanced deep-learning architectures such as TC-LSTM and bidirectional LSTMs. Our analysis, spanning from April 15^{th}, 2020 to April 15^{th}, 2022, shows that FIGI-Net is an improvement over state-of-the-art models, successfully predicting COVID-19 dynamics at county, state, and national level, across multiple time horizons, reaching error reductions of up to 78.63%, 85.85% and 86.48% in RRMSE, accordingly.

At county level, FIGI-Net successfully predicted COVID-19 activity, scoring error reductions of up to 87.38% in comparison to the baseline model, Persistence. FIGI-Net consistently placed as a top 1 performer across the multiple time horizons based on the RMSE and RMSE metrics, as presented in Table 1. At state level, we compared FIGI-Net predictions against state-of-the-art models for the 1-day, 7-day and 14-day horizon. Our results showed that FIGI-Net scored error reductions in comparison to Persistence model of 87.99%, 56.01%, and 41.15% in RMSE, accordingly (see Table 2). Finally, at national level, our model successfully presented an error reduction of up to 86.48%. We attribute the success of FIGI-Net across multiple geographical resolutions due to the pre-trained model component in our framework, which captures meaningful patterns across global infection dynamics. This clustering-based approach utilizes global spatio-temporal features learnt a priori, enabling submodels to mitigate the influence of irrelevant information and increase its own predictive power. This innovative framework addresses the issue of rapidly changing dynamics (see, for example, Figure 2), ensuring accurate forecasting within short-term periods, up to 2 weeks.

The analysis of FIGI-Net predictive power against with the CDC's official ensemble predictions showed a notable improvement as our model showcased a substantial reduction in error rates. At the county level, our model demonstrated an impressive more than 50% decrease in error, while at the state level, the reduction stood at least 13% (see Table 4 and 5). This success is indicative of our model's adaptation capacity, particularly at the granular county level—a domain where the CDC's ensemble predictions exhibited comparatively poorer performance. This suggests the potential of our model not only in refining county-level forecasts, but also in addressing the nuances that contribute to more accurate forecasting, highlighting its utility in augmenting current predictive methodologies.

The forecasts from FIGI-net presented in this work were created using a training moving window consisting of 75 days of data in length. Our choice of training window is based on an experimental analysis on the predictive power of FIGI-net as a function of the training window size. Figure 7 illustrates that longer length of training data positively influenced our model's performance. Particularly, there is a small change in 2-week horizon's performance when the training data length exceeded 75 days, whereas a longer training data length shows enhanced performance for a 1 week horizon. Based on our 1 and 2 week horizon's performances (presented in Figure 4), we determined the optimal training data length to be 75 days. Moreover, when comparing our predictions against the CDC's ensemble, we observed that our model outperforms with exceeding 60 days of training data.These results indicate that while longer training data enhances the our model's accuracy in predicting infection numbers over a 1 week horizon, there is limited improvement in the 2 week horizon when the training data length

Table 6: Performance of different forecasting models during critical time periods in 1-week ahead

Model	Persistence	FIGI-Net	CDC	Karlen	CovidAnalysis	JHU	MIT	Microsoft
Slope Dissimilarity	0.3121±0.0792	**0.1365**±0.0755	0.2909±0.1067	0.5173±0.1463	0.5626±0.1361	0.2742±0.1344	0.5809±0.1638	0.45±0.1336
RMSE	11367.35±14539.1296	**5934.76**±6393.4828	11175.61±13640.4815	15684.27±14568.3393	17112.11±20400.6492	9911.12±9031.9252	17459.79±22030.9831	20079.54±25511.9738
MAPE	20.89±3.5791	**11.1**±2.9743	30.05±26.944	66.06±74.1738	51.52±23.2128	34.51±55.9652	47.46±29.0615	41.11±47.1091
P-value	<0.001	-	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 7: Performance of different forecasting models during critical time periods in 2-week ahead

Model	Persistence	FIGI-Net	CDC	Karlen	CovidAnalysis	JHU	MIT	Microsoft
Slope Dissimilarity	0.3121±0.0792	**0.1995**±0.0932	0.4288±0.1151	0.5155±0.1467	0.6181±0.1372	0.4933±0.1198	0.746±0.1812	0.6213±0.1693
RMSE	18314.08±24534.4306	**7624.24**±8866.1508	16953.5±21673.4287	17330.03±17305.0527	23141.39±29330.5663	9424±9912.5806	29344.63±39638.1569	28787.66±36694.9282
MAPE	30.84±3.9001	**13.8**±3.8517	37.05±19.4212	70.71±77.0026	60.56±20.7622	38.98±50.6652	61.14±24.0137	49.75±44.2705
P-value	<0.001	-	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

exceeds 75 days. Overall, our model demonstrates improved infection prediction with a minimum training data length of 2 months

Upon analyzing the geographical distribution maps presented in Figure 5, which illustrate our county-level predictions and errors across the US, it is evident that variations in reporting values might arise from changes in epidemic prevention policies in different regions. These differences in data format can significantly complicate the accurate prediction of COVID-19 infections and lead to a substantial increase in errors. For example, during the Delta-variant wave, as shown in Figure 5(d)-(f), the RMSE for counties in Kansas exceeded that of neighboring states, even though the pandemic risk in those areas was relatively mild. This pattern is also noticeable in Iowa during the Omicron-variant wave, as depicted in Figure 5(g)-(i). We attribute these observations to two factors: (1) the relationship between low daily reported infection cases and higher predicted outcomes, leading to larger errors. For instance, if a county reports 2 infection cases while the prediction is 4 cases, this discrepancy results in a larger RMSE. (2) State governments change their recording policies from daily to weekly at certain periods, introducing inconsistencies and irregularities in the data format that could impact model predictions. This policy change points out the importance of consistent instructions and practices for specific epidemiological diseases, ensuring effective management of public healthcare information and promoting accurate disease analysis, prediction, and prevention. Further studies are necessary to explore and address these issues in order to enhance the accuracy and reliability of infection predictions.

As depicted in Figure 6.A and B, most tested forecasting models struggled to predict the direction of future infection trends accurately, primarily due to the highly variable and fast transmission rate of COVID-19. However, our FIGI-Net model accurately predicted the trend direction during the initial stage of each outbreak. This capability is attributed to our model's daily prediction of infection case numbers, which has allowed for the early detection and response to sudden changes. Furthermore, utilizing county-level data with clustering allows the identification of early regional variations and swift adjustment of the forecasting trend by the proposed sub-models. These features enable our model to efficiently adapt to dynamically various changes in infection numbers and trends during COVID-19 outbreaks. Moreover, our model exhibited a higher slope similarity score (See Table 6 and 7), lower RMSE and MAPE scores than others. These results indicate that the proposed model excels at predicting the direction of forecasting trend, facilitating early implementation of COVID-19 transmission prevention. Our study underscores the robustness and effectiveness of time-series deep learning-based methods in handling dynamic and sudden changes in infection numbers during short-term time periods.

Based on the experiments, our proposed model has certain limitations. Firstly, our deep learning-based model requires a longer training time, compared to linear models, due to the complexity and computational demands. Although our model can automatically obtain optimal hyperparameters, this leads to extended convergence times. Additionally, each cluster has its trained model to enhance forecasting outcomes, but this increases computational time for predicting a single time period. Another limitation is the requirement for an adequate amount of training data. Deep learning models need large volumes of diverse and representative data to learn underlying patterns effectively and make accurate predictions. Limited data can compromise the model's performance. Therefore, ensuring a substantial amount of high-quality training data is essential for our model's effectiveness. This could be achieved by increasing the granularity of infection case data, such as building town-level datasets. Addressing these limitations is crucial for the model's real-world application. Future research aims to overcome these challenges and enhance the model's performance under various conditions and tasks.

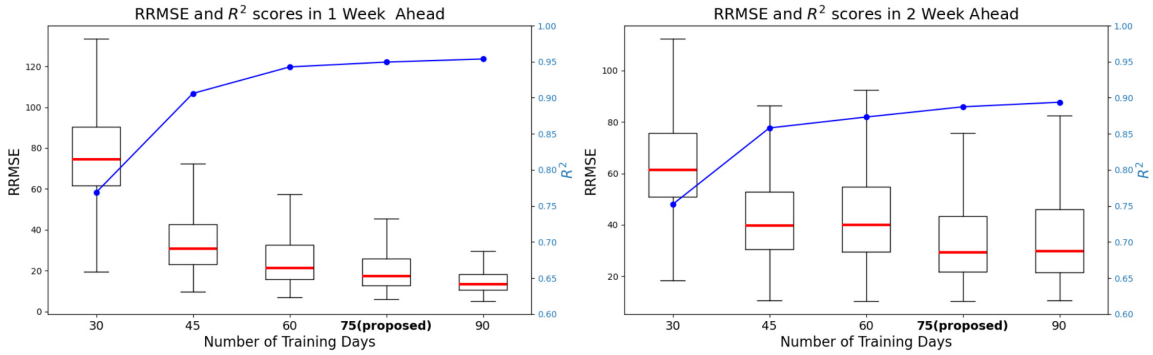

Figure 7: Comparison of training day lengths for the proposed model in 1 and 2 weekly horizons. The figure shows the RRMSE values and R^2 scores (represented by the blue line) achieved by the proposed model using different training day lengths. For the 1 weekly horizon, 75-day and 90-day training lengths result in lower RRMSE prediction errors, with a median value below 20 and an approximate R^2 score of 0.95. Moreover, increasing the training day length beyond 75 days leads to a 26.3% reduction in RRMSE for 2 weekly ahead prediction but does not further improve performance with longer training periods. These findings highlight the significant impact of training day length on short-term infection forecasting.

In conclusion, the FIGI-Net model represents an improvement in the field of COVID-19 infection forecasting. By employing temporal clustering and a stacked structure of biLSTM, our model achieves accurate and efficient COVID-19 infection forecasts from fine-grained county level datasets. Accurate and early predictions of COVID-19 outbreaks at the county, state, and federal levels is of paramount importance for effective public health management. Our model's ability to provide early warning of potential outbreaks allows prompt and targeted public health interventions. The potential applications of our model in public health management and epidemiological disease prevention are substantial and could profoundly impact mitigating the effects of future infectious disease outbreaks.

Data and Methods

Data Collection and Cleaning

The data utilized in this study includes the daily COVID-19 cumulative infectious and death cases of US counties, obtained from the Johns Hopkins Center for Systems Science and Engineering (CSSE) Coronavirus Resource Center between January 21st, 2020 and April 16th, 2022[14]. It is important to note that each county or state government may have different policies for pandemic recording and reporting, which can make the CSSE data difficult to evaluate and analyze. Additionally, cumulative data may not efficiently differentiate regional variation. To address these issues, we computed utilized a 7-day average method to denoise the daily official COVID-19 cases [30]. Because CSSE data occasionally had abnormal or missing-data observations, we selected valid data of US counties, including 3143 counties between February 2020 to March 2022, as the ground truth for further evaluation. We then separated this time-series data into 48 infection forecasting time periods, each of which is approximately 15 days in length and utilized this data to train our models.

Temporal Clustering

Based on the evidence that neighboring COVID-19 dynamics may highly influence local dynamics (see Noor *et al* [38]), our methodology incorporates a two-step spatio-temporal clustering procedure that aims to identify both global and local similarities between the COVID-19 activity within each county in the US. The outcome of this analysis guides our framework to train sub-models that are fit only on the most relevant dynamics for each county. The procedure consists in the following steps: 1) Creating a set of feature vectors representing the similarity between each county, 2) Compressing the representation of this features via a dimensionality-reduction procedure (in this case, we apply

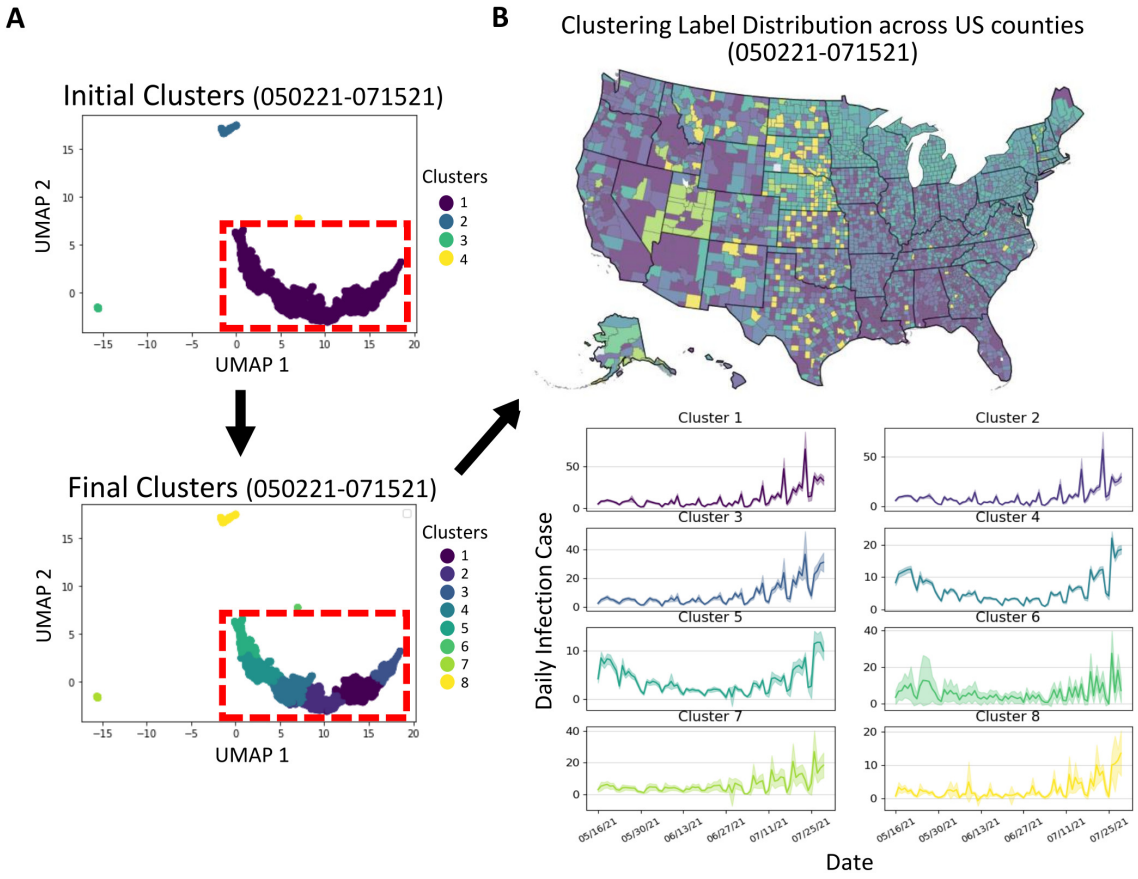

Figure 8: An example of temporal clustering in a specific time period. (A) The UMAP method is utilized to map the correlated temporal features, which are then grouped to form initial clusters. The largest cluster (highlighted by the red-dashed rectangle) is further subdivided to obtain subclusters, and a total of 8 clusters are reordered based on the average infection number of the last 7 days within each cluster. (B) The geographical distribution across US counties exhibits how temporal clustering captures the relationship between time-series and spatial information. The infection curves of the training data (with a duration of 75 days) for the clusters, along with a 95% confidence interval, illustrate that this approach provides a collection of highly related yet distinct subclusters to help the sub-model efficiently learn and make accurate predictions.

UMAP), and 3) A 2-level clustering procedure over the resulting lower dimensional version of these feature vectors.

Creating the feature vectors For each time period, we computed the autocorrelation of daily COVID-19 confirmed cases and fatality rates, and obtained the cross-correlation between them. These two correlation features were concatenated per county, resulting in feature vectors representing the combined correlation information for each county.

Dimensionality reduction Given our feature vectors are of dimension 301 given the number of counties within the us, a lighter representation of this feature vectors is necessary. We transformed our feature vectors using the UMAP method [34]. UMAP is dimensionality reduction technique to effectively prevent and handle global and local nonlinear structure in a lower-dimensional space. We selected UMAP as our dimensionality-reduction step given its ability to make an optimal choice for preserving both local and global relationships during the reduction process. The resulting outcome in this step is a feature vector of dimension 2.

2-level clustering To identify temporal clusters, we used the unsupervised DBScan method [17] to extract initial clusters, followed by selecting the largest cluster to obtain the subclusters via spectral clustering [36]. Figure 8 shows a representative example illustrating the process of temporal clustering. The initial clustering identifies the global differences in county trends, and the subclustering

distinguishes local differences from similar reporting trends. By seeking similar temporal features, we identified the clusters that include counties from different geographical locations, enabling the provision of relevant local information. Subsequently, the list of cluster labels was rearranged in descending order of infection risk based on the average infection count over last 7 days within each cluster. According to the results of our experiments, we determined the optimal number of clusters to be 8 for facilitating further submodel training. When the number of clusters was too small, the submodel may learn irrelevant infection features that could adversely impact local predictions. Conversely, if we identified too many clusters, the size of training data becomes smaller and may not provide enough training data for submodel training. We found that 8 clusters achieve a good balance between providing relevant local information for accurate predictions and ensuring an adequate amount of training data.

Training sub-models through transfer learning from a global model

Due to the dynamic and rapid changes in the trend of COVID-19, the use of time series forecasting models has become essential. The LSTM model, developed from the recurrent neural network, is particularly useful in handling time series forecasting problems [23]. In order to address the short-term infection variability, we implemented a bidirectional stacked LSTM (biLSTM) model for predicting infection case trends, and the architecture of the proposed model is shown in Figure 1(b). It leverages the bidirectional method to learn the variability of the future sequence trends over time and strengthens the ability to handle unexpected sudden changes. Furthermore, the stacked structure helps our model recognize the similarity and relevance of the entire historical trend of each time period among counties.

To achieve accurate and efficient forecasts in each time period, we introduced transfer learning to deal with the rapid variability and transmission of COVID-19. We collected 75-day length raw data of all counties before the time period for forecasting and used a 60-day length sliding window, consisting of 45-day length for training inputs and 15-day length for predicted labels, to generate the training dataset of all counties and the counties in each temporal cluster, respectively. We used the training dataset to train the global model to learn universal infection features to address possible trend changes. Then we transferred the parameters of this pre-trained global model to train 8 submodels for each temporal cluster, and each sub-model was fine-tuned by using the county data of each temporal cluster. During each model training, the training data was randomly split into training (80%) and validation sets (20%). The temporal clustering provides highly relevant historical information, enabling the training of local submodels by optimizing the parameters of the global pre-trained model. This approach ensures accurate forecasting of county-level infection numbers. We used the Adam optimization algorithm, with a learning rate of 10^{-3} and 100 epochs during the training process.

Model evaluation

Daily forecasting error is assessed using the RMSE and RRMSE, defined as:

$$RMSE = \sqrt{\sum \frac{(y-\hat{y})^2}{N}} \times 100$$

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum (y-\hat{y})^2}}{\sqrt{\frac{1}{N}\sum y^2}} \times 100$$

,where y, $\hat{y}$, and N represent the observation, the predicted values, and the number of US counties, respectively. Also, to assess the consistency of the model over time, we use MAPE, defined as:

$$MAPE = \frac{1}{N}\sum \left|\frac{y-\hat{y}}{y}\right| \times 100.$$

We also evaluated the similarity of slope score, which indicate the difference of trend directions between observations and forecasting models, to measure the accuracy of predicted directions in each time point. The range of slope score is from 0 to 1 and the function is defined as follows,

$$Slope\ Difference = \frac{2 \times \left|\tan^{-1}(Slope\_Observation) - \tan^{-1}(Slope\_Prediction)\right|}{\pi}$$

A lower score represents a predicted trend direction that aligns with the ground truth at a specific time point. We employed the linear regression method [18] to calculate the slope of each point by using short time interval, which includes the two points before and after the specific point. These measurement methods were used to assess the accuracy of the forecasting model in reflecting rapid trend changes. Additionally, the two-sided Wilcoxon rank sum test, also known as the Mann-Whitney U test, was utilized to test statistical significance among the model performances [37]. This testing approach is frequently employed for handling performance metrics without assuming a normal distribution of the data and without specifying the direction of the difference.

Data Availability Statement

The data used in this study are publicly available and consist of daily COVID-19 cumulative infectious and death cases reported for U.S. counties. The dataset was obtained from the Johns Hopkins Center for Systems Science and Engineering (CSSE) Coronavirus Resource Center, spanning from January 21st, 2020, to April 16th, 2022 [14]. The dataset can be directly accessed from the Johns Hopkins CSSE Coronavirus Resource Center website (https://github.com/CSSEGISandData/COVID-19). Researchers interested in utilizing the data for further analysis can refer to the original source for detailed documentation on data collection methods and definitions. For additional information or inquiries about the dataset, please visit the website or contact the Johns Hopkins CSSE Coronavirus Resource Center.

Acknowledgement

This work was supported by NIH, United States (Grant Number: R35GM133725).

References

[1] Cdc, centers for disease control and prevention, mmwr weeks). `https://stacks.cdc.gov/view/cdc/22305`. Accessed: 2022-11-15.

[2] Roy M Anderson, Hans Heesterbeek, Don Klinkenberg, and T Déirdre Hollingsworth. How will country-based mitigation measures influence the course of the covid-19 epidemic? *The lancet*, 395(10228):931–934, 2020.

[3] Sina F Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M Atkinson. Covid-19 outbreak prediction with machine learning. *Algorithms*, 13(10):249, 2020.

[4] Joan L Aron and Ira B Schwartz. Seasonality and period-doubling bifurcations in an epidemic model. *Journal of theoretical biology*, 110(4):665–679, 1984.

[5] KE ArunKumar, Dinesh V Kalaga, Ch Mohan Sai Kumar, Masahiro Kawaji, and Timothy M Brenza. Comparative analysis of gated recurrent units (gru), long short-term memory (lstm) cells, autoregressive integrated moving average (arima), seasonal autoregressive integrated moving average (sarima) for forecasting covid-19 trends. *Alexandria Engineering Journal*, 61(10):7585–7603, 2022.

[6] Soumyanil Banerjee, Ming Dong, and Weisong Shi. Spatial–temporal synchronous graph transformer network (stsgt) for covid-19 forecasting. *Smart Health*, 26:100348, 2022.

[7] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[8] Logan C Brooks, David C Farrow, Sangwon Hyun, Ryan J Tibshirani, and Roni Rosenfeld. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS computational biology*, 14(6):e1006134, 2018.

[9] Kaustuv Chatterjee, Kaushik Chatterjee, Arun Kumar, and Subramanian Shankar. Healthcare impact of covid-19 epidemic in india: A stochastic mathematical model. *Medical Journal Armed Forces India*, 76(2):147–155, 2020.

[10] Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489):395–400, 2020.

[11] Estee Y Cramer, Yuxin Huang, Yijin Wang, Evan L Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul H Kanji, Ayush Khandelwal, Khoa Le, Jarad Niemi, Ariane Stark, Apurv Shah, Nutcha Wattanachit, Martha W Zorn, Nicholas G Reich, and US COVID-19 Forecast Hub Consortium. The united states covid-19 forecast hub dataset. *medRxiv*, 2021.

[12] Domenico Cucinotta and Maurizio Vanelli. Who declares covid-19 a pandemic. *Acta Biomedicine*, 91(1):157–160, 2020.

[13] Jonas Dehning, Johannes Zierenberg, F Paul Spitzner, Michael Wibral, Joao Pinheiro Neto, Michael Wilczek, and Viola Priesemann. Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science*, 369(6500):eabb9789, 2020.

[14] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[15] Shawni Dutta and Samir Kumar Bandyopadhyay. Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release. *MedRxiv*, 2020.

[16] Martin Enserink and Kai Kupferschmidt. With covid-19, modeling takes on life and death importance, 2020.

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[18] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[19] James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.

[20] Robert Harrison. Introduction to monte carlo simulation. *AIP conference proceedings*, 1204:17–21, 01 2010.

[21] Andres Hernandez-Matamoros, Hamido Fujita, Toshitaka Hayashi, and Hector Perez-Meana. Forecasting of covid19 per regions using arima models and polynomial functions. *Applied soft computing*, 96:106610, 2020.

[22] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[24] Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O'Banion. Examining covid-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113*, 2020.

[25] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[26] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*, 20(5):553–558, 2020.

[27] Dezhi Li, Guanying Huang, Guomin Zhang, and Jiangbo Wang. Driving factors of total carbon emissions from the construction industry in jiangsu province, china. *Journal of Cleaner Production*, 276:123179, 2020.

[28] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490):489–493, 2020.

[29] Jose Lourenco, Robert Paton, Mahan Ghafari, Moritz Kraemer, Craig Thompson, Peter Simmonds, Paul Klenerman, and Sunetra Gupta. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the sars-cov-2 epidemic. *MedRxiv*, pages 2020–03, 2020.

[30] Fred S Lu, Mohammad W Hattab, Cesar Leonardo Clemente, Matthew Biggerstaff, and Mauricio Santillana. Improved state-level influenza nowcasting in the united states leveraging internet-based data and network approaches. *Nature communications*, 10(1):1–10, 2019.

[31] Fred S Lu, Andre T Nguyen, Nicholas B Link, Mathieu Molina, Jessica T Davis, Matteo Chinazzi, Xinyue Xiong, Alessandro Vespignani, Marc Lipsitch, and Mauricio Santillana. Estimating the cumulative incidence of covid-19 in the united states using influenza surveillance, virologic testing, and mortality data: Four complementary approaches. *PLOS Computational Biology*, 17(6):e1008994, 2021.

[32] Ruifang Ma, Xinqi Zheng, Peipei Wang, Haiyan Liu, and Chunxiao Zhang. The prediction and analysis of covid-19 epidemic trend by combining lstm and markov method. *Scientific Reports*, 11(1):1–14, 2021.

[33] Benjamin F Maier and Dirk Brockmann. Effective containment explains subexponential growth in recent confirmed covid-19 cases in china. *Science*, 368(6492):742–746, 2020.

[34] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[35] Oriol Mitjà, Àlex Arenas, Xavier Rodó, Aurelio Tobias, Joe Brew, and José M Benlloch. Experts' request to the spanish government: move spain towards complete lockdown. *The Lancet*, 395(10231):1193–1194, 2020.

[36] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. NIPS'01, page 849–856. MIT Press, 2001.

[37] Magdalena Niewiadomska-Bugaj and Robert Bartoszynski. *Probability and statistical inference*. John Wiley & Sons, 2020.

[38] Aziz Ullah Noor, Farhana Maqbool, Zulfiqar A Bhatti, and Asmat Ullah Khan. Epidemiology of covid-19 pandemic: Recovery and mortality ratio around the globe. *Pakistan journal of medical sciences*, 36(COVID19-S4):S79, 2020.

[39] Sen Pei and Jeffrey Shaman. Initial simulation of sars-cov2 spread and intervention effects in the continental us. *MedRxiv*, 2020.

[40] Evan L Ray, Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana, Xinyue Xiong, et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRXiv*, pages 2020–08, 2020.

[41] Jeffrey Shaman and Alicia Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[42] R at al Sujath, Jyotir Moy Chatterjee, and Aboul Ella Hassanien. A machine learning forecasting model for covid-19 pandemic in india. *Stochastic Environmental Research and Risk Assessment*, 34(7):959–972, 2020.

[43] Loring J Thomas, Peng Huang, Fan Yin, Xiaoshuang Iris Luo, Zack W Almquist, John R Hipp, and Carter T Butts. Spatial heterogeneity can lead to substantial local variations in covid-19 timing and severity. *Proceedings of the National Academy of Sciences*, 117(39):24180–24187, 2020.

[44] Wan Yang, Jaimie Shaff, and Jeffrey Shaman. Covid-19 transmission dynamics and effectiveness of public health interventions in new york city during the 2020 spring pandemic wave. *medRxiv*, 2020.