npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01606-1

Fine-grained forecasting of COVID-19 trends at the county level in the United States

Check for updates

Tzu-Hsi Song[1], Leonardo Clemente[2], Xiang Pan[1], Junbong Jang[1], Mauricio Santillana[2] ✉ & Kwonmoo Lee[1] ✉

The novel coronavirus (COVID-19) pandemic has had a devastating global impact, profoundly affecting daily life, healthcare systems, and public health infrastructure. Despite the availability of treatments and vaccines, hospitalizations and deaths continue. Real-time surveillance of infection trends supports resource allocation and mitigation strategies, but reliable forecasting remains a challenge. While deep learning has advanced time-series forecasting, its effectiveness relies on large datasets, a significant obstacle given the pandemic's evolving nature. Most models use national or state-level data, limiting both dataset size and the granularity of insights. To address this, we propose the Fine-Grained Infection Forecast Network (FIGI-Net), a stacked bidirectional LSTM structure designed to leverage county-level data to produce daily forecasts up to two weeks in advance. FIGI-Net outperforms existing models, accurately predicting sudden changes such as new outbreaks or peaks, a capability many state-of-the-art models lack. This approach could enhance public health responses and outbreak preparedness.

Since 2019, the SARS-CoV-2 coronavirus has spread in human populations and causing a disease called COVID-19. Its rapid spread, from several countries to the global stage, prompted the World Health Organization (WHO) to declare it a global pandemic in early 2020[1]. With over half a billion infections and more than 6 million deaths recorded worldwide[2], COVID-19 has significantly reshaped society, national healthcare systems, and the global economy. Due to its fast mutation rate, new outbreaks of COVID-19 have continued 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[3].

Epidemic forecasting in this context has presented 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[4,5]. 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[6–19]. However, spatial and temporal heterogeneity known to exist across locations, such as socioeconomic and demographic factors, the diverse implementation of local mitigation policies, and the emergence of new variants, are often very challenging to incorporate and appropriately update in these models. As a result, these methods often fall short in capturing essential local heterogeneities in infection dynamics[20,21].

In contrast, data-driven methods may excel in this context by implicitly incorporating the influence of diverse local policies, new variants, demography, and socioeconomic factors, by learning directly from the reported data. 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[22].

Data-driven machine learning models have been extensively applied in disease forecasting[23–25]. For instance, Sujath et al. conducted a comparison of linear regression, multilayer perceptron (MLP), and vector autoregression, with the MLP model demonstrating superior performance[26]. Ardabili et al. integrated genetic algorithms with SIR and SEIR models to predict outbreak trends[27]. Hernandez et al. utilized an Auto-regressive Integrated Moving Average (ARIMA)[28] model with polynomial functions for global COVID

[1]Vascular Biology Program and Department of Surgery, Boston Children's Hospital, Harvard Medical School, Boston, MA, 02115, USA. [2]Department of Physics and Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, 02115, USA. ✉e-mail: m.santillana@northeastern.edu; kwonmoo.lee@childrens.harvard.edu

infection trend predictions, using cumulative datasets[29]. Additionally, Lu et al. compared and integrated various approaches, combining statistical linear and nonlinear models to estimate the cumulative number of weekly confirmed cases[30]. Many of these models rely on assumptions about the real-time availability of reliable data, a stable non-evolving pathogen, and/or population-level adherence to specific behaviors during epidemic outbreaks. However, the dynamics of COVID-19 have been in constant flux, influenced by factors such as virus characteristics and heterogeneous population adherence to mitigation policies such as recommendations to stay at home or mask wearing[31]. 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, especially after enough observations to train models have been recorded. 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 their models to adapt to sparse data situations. Bandyopadhyay et al. used LSTM with a gate circulation network to estimate COVID-19 cases[32], while Huang et al. employed CNNs to predict cumulative COVID-19 deaths[33]. Ruifang et al. combined LSTM with a Markov method for national cumulative COVID-19 predictions[34]. 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[35]. Additionally, Transformer models, known for their proficiency in natural language processing, have been utilized in COVID-19 forecasting. Soumyanil et al. developed a graph transformer network with synchronous temporal and spatial information[36], while Kapoor et al. used a similar approach to predict COVID-19 confirmed cases[37]. However, the successful application of deep learning depends on the availability of large datasets, a requirement that presents a significant challenge in the context of the COVID-19 pandemic characterized by rapidly evolving epidemiological conditions. These deep learning models have relied on national or state-level data, inherently limiting the dataset size and causing the local heterogeneity of COVID-19 infections to be averaged out. Consequently, despite the impressive capabilities of deep learning, its forecasting results were compromised, 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 awareness 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), utilizes fine-grained time series of COVID-19 infection data at the county level in the U.S. This approach leverages LSTM capabilities with a stacked bidirectional component, enhancing the model's capacity to discern diverse global infection trends. Subsequently, upon identifying clusters of counties with similar COVID-19 temporal patterns, we applied transfer learning from the global biLSTM model trained on segmented historical daily confirmed cases of U.S. counties, and trained specific cluster-based biLSTM models. This cluster transfer learning concept preserves the model's inherent capabilities[38–42] while empowering it to swiftly adapt to short-term trend changes locally, refining forecasting accuracy. Using our proposed FIGI-Net, we conducted predictions for future COVID-19 infection cases, encompassing both short-term and mid-term forecasts spanning from the next day to the next 15 days. The key contributions of our framework are summarized as follows:

- We provide a practical approach to determine the minimum amount of observations –length of the initial training time period – needed for deep learning-based machine learning models to deliver reliable daily forecasts of disease activities in the context of a novel and emerging disease outbreak.
- 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 infectious clusters of COVID-19 temporal data. The structure of the proposed framework is presented in Fig. 1 and additional details are provided in the section "Training sub-models through transfer learning from a global model".
- Given the explicit need to produce a model responsive and adaptive to dynamic changes of disease transmission due to a diverse set of factors – changes in human behavior, availability of vaccines, and testing capabilities, our model is dynamically re-trained on a moving-window that only uses the most recent trends as input. This time window is chosen prior to evaluating our model's forecasts in a strictly out-of-sample fashion.
- Focusing on U.S. county-level COVID-19 time series data, we show FIGI-Net's efficiency and accuracy in forecasting sharp changes in COVID-19 activity, both in short-term (up to 1 day) and mid-term (up to 2 weeks) forecasting scenarios.
- Furthermore, our methodology is extended to national and state-level forecasts using weekly time periods. The results highlight the

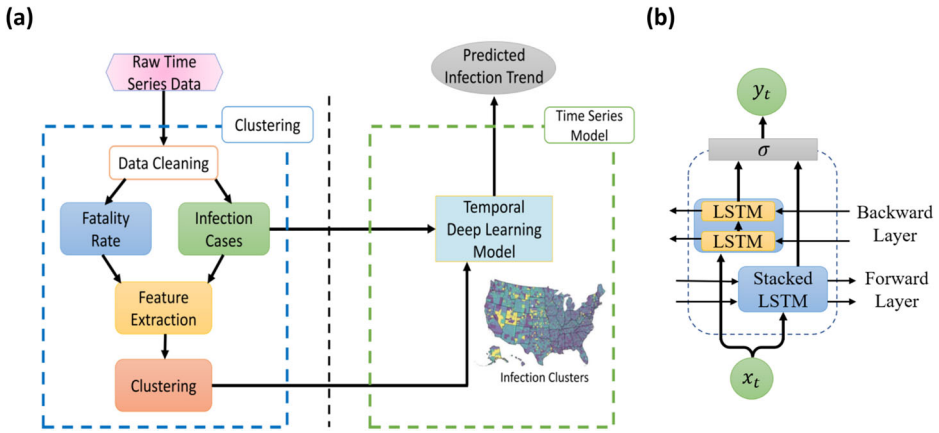


Fig. 1 | Our proposed model, FIGI-Net, architecture for COVID-19 infection prediction. a A diagram that illustrates our methodology's framework, consisting of a clustering and a time series deep learning component. In the clustering section, auto-correlation and cross-correlation were used to extract features from the pre-cleaning infection cases and fatality rate data. Then, we applied a clustering technique to the extracted features to identify similar COVID-19 dynamics. The temporal deep learning model utilized the identified 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, enabling it to effectively handle short-term state changes. The σ function represents the merging function.

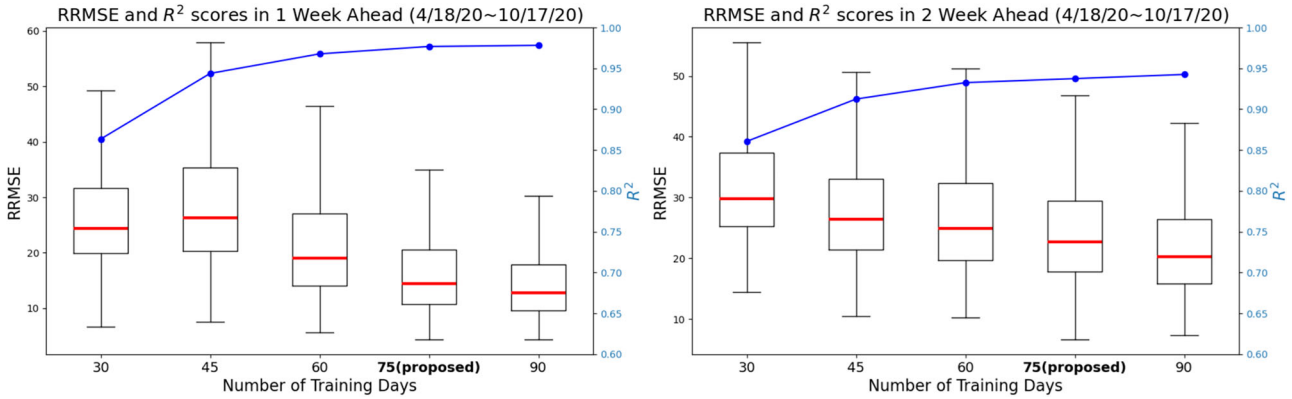


Fig. 2 | Comparison of training day length identification for the proposed model in 1 and 2 weekly horizons. The figure shows the RRMSE values and R^2 scores (represented by the blue line) achieved by the proposed model using different training day lengths from first 6-month dataset (from April to Oct, 2020). For the 1-week 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-week 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.

superiority of our FIGI-Net, showcasing more than a 40% error reduction 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 October 18th, 2020 and April 15th, 2022. These models included the model we propose: FIGI-Net, as well as Autoregressive statistical models, a collection of recurrent neural network (RNN) based models (GRU, LSTM), a stacked bidirectional LSTM (biLSTM), a set of LSTM-based models incorporating temporal clustering (TC-LSTM and TC-biLSTM), and a "naive" (Persistence) model[43] to be used as a baseline. Additionally, for comparison purposes, we implemented five additional models commonly found in the forecasting literature: ARIMA[28], Prophet[44], and Transformer-based models (Transformer[45], Informer[46], and Autoformer[47]) both with and without clustering.

We used our models to retrospectively produce daily forecasts for multiple time horizons, h, ranging from $h = 1, 2. . . , 15$ days ahead. Visual representations of our predictions, alongside the actual observed COVID-19 confirmed cases, are presented in Figs. 3, 4, and 5 for county, state and national levels. We evaluated our model's performance by comparing our out-of-sample predictions with subsequently observed reported data in each time horizon h using multiple error metrics that include: the root mean square error (RMSE), relative RMSE (RRMSE), and mean absolute percentage error (MAPE) metrics, as detailed in the section "Model evaluation". In addition, we compared the prediction performance of our model with a diverse set of statistical and machine learning models, including state-of-the-art forecasting methodologies reported by the CDC (a comprehensive list of these models is provided in the section "Comparative analysis of COVID-19 forecasting models during critical time periods"). Finally, we assessed our model's ability to anticipate the onset of multiple outbreaks during our studied time periods.

Determining the length of moving time windows for training

Given that neural network based models typically need a large amount of data to be trained[48,49], we first investigated the minimum amount of data that would be necessary for our model to produce robust and reliable forecasts. We investigated the length of the moving window for training necessary to yield forecasts responsive to changes in disease dynamics due to changes in human behavior over time, vaccine availability, different transmission intensities for different variants (e.g. omicron), among other factors. Using observations from April 18th, 2020 to October 17th, 2020 we identified a time window size of 75 days to be a good compromise between reliable forecasting performance (shown in Fig. 2) and a short enough time period that would allow the model to continuously learn new transmission patterns. For more details, please refer to the Discussion.

FIGI-Net forecasting performance at the county level

For each forecasting horizon h, we computed FIGI-Net's prediction error metrics (RMSE and RRMSE) across all counties, over the time period: 10/18/2020 - 4/15/2022. Table 1 shows the prediction error values and percentage of error reduction with respect to the Persistence model of all the models for 1-day, 7-day and 14-day horizons. Based on the experiments regarding training length influence, we evaluated all comparative models using an optimal training length of 75 days in the following eighteen months of data. Our results demonstrate that FIGI-Net model has the greatest error reduction across all horizons (90%, 83%, and 35% RMSE reduction, correspondingly), followed by the biLSTM model (85% in horizon 1-day, 75% in horizon 7-day, and 34% in horizon 14-day RMSE reduction). To assess the significance of these error reductions, we performed the two-sided Wilcoxon signed rank test[50] over the entire outcomes of tasks.

In Fig. 3(a), we visualize the forecasting ability of each model for both the 1-day and 7-day ahead tasks, including classic models such as Persistence (a naive rule stating $y_{t+1} = y_t$)[43], Autoregressive models(AR)[51], ARIMA[28], and Prophet, as well as deep learning-based models like GRU or LSTM architectures, and Transformer-based models such as Transformer, Autoformer, and Informer. 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 (RRMSE) scores across all models for both the 1-day ahead and the 7-day ahead prediction task (approximately 6.98% of RRMSE at 1-day ahead, and 9.92% at 7-day ahead), followed by deep learning models with bidirectional components (TC-biLSTM with 17.08% and 23.03% of RRMSE and biLSTM with 9.83% and 14.28% of RRMSE, respectively). All recurrent neural network-based models showed improvement over the Persistence model (83.13% and 50.48% of RRMSE), the Autoreggressive model (81.57% and 83.87% of RRMSE), ARIMA (86.68% and 77.13% of RRMSE), and Prophet (64.3% and 71.91% of RRMSE). In Fig. 3(b), FIGI-Net reduced the errors by approximately 90% in the 1-day ahead prediction and by around 83% in the 7-day ahead prediction compared to the Persistence model. Moreover, Transformer-based models exhibit higher RRMSE values in the 1-day, 7-day, and 14-day ahead horizons compared to the recurrent neural network-based models (also shown in Fig. 3(a) and (b)). Transformer models typically require large amounts of data to perform optimally. Consequently, they underperformed in COVID-19 forecasting due to limitations in data

Table 1 | Performance metrics with error reduction for the 1-day, 7-day and 14-day ahead tasks at the county level

Model	RMSE			RRMSE(%)			Error Reduction		
	1st day	7th day	14th day	1st day	7th day	14th day	1st day	7th day	14th day
Persistence	45.11 ± 176.33	29.71 ± 133.62	36.56 ± 152.9	83.13 ± 111.19	50.48 ± 48.18	65.48 ± 69.41	–	–	–
Autoregressive	44.43 ± 209.3	49.15 ± 9.68×10^{4}	65.91 ± 10^{9}	81.57 ± 95.71	83.87 ± 3.18×10^{4}	103.16 ± 8.72×10^{8}	1.01 ± 1.13	1.5 ± 702.98	1.54 ± 1.12×10^{7}
ARIMA	51.35 ± 324.35	44.84 ± 306.39	38.58 ± 1.2×10^{4}	86.68 ± 782.34	77.13 ± 470.71	68.95 ± 4.2×10^{4}	1.06 ± 8.46	1.39 ± 12.24	0.99 ±;649.78
Prophet	33.93 ± 121.2	39.84 ± 141.58	45.27 ± 170.04	64.3 ± 33.46	71.91 ± 47.53	81.28 ± 93.05	0.81;± 0.38	1.28;± 0.84	1.21 ±;0.77
GRU	18.62 ± 101.65	19.23 ± 106.35	23.75 ± 118.99	33.26 ± 24.04	34.45 ± 25.11	43.61 ± 25.27	0.49 ± 0.3	0.68 ± 0.4	0.7 ± 0.33
LSTM	18.44 ± 91.71	18.01 ± 104.53	22.83 ± 118.66	31.29 ± 24.09	33.37 ± 25.08	42.04 ± 24.9	0.46 ± 0.3	0.64 ± 0.4	0.7 ±;0.33
TC-LSTM	19.03 ± 105.94	19.22 ± 112.28	23.56 ± 123.17	33.19 ± 23.1	35.09 ± 26.23	43.84 ± 25.97	0.49 ± 0.3	0.7 ± 0.4	0.72 ± 0.34
TC-biLSTM	10.47 ± 59.91	12.96 ± 80.49	23.66 ± 115.85	17.08 ± 20.25	23.03 ±;30.25	44.33 ± 37.06	0.25 ± 0.24	0.43 ± 0.51	0.73 ± 0.4
Transformer	31.53 ± 94.37	33.37 ± 113.29	37.44 ± 128.22	57 ± 34.43	61.31 ± 33.67	68.47 ± 40.67	0.75 ± 0.38	1.04 ± 0.75	0.97 ± 0.58
Informer	33.04 ± 118.27	35.73 ± 122.35	41.1 ± 134.62	61.78 ± 31.94	63.63 ± 35.47	70.34 ± 49.4	0.77 ± 0.34	1.08 ± 0.78	1.02 ± 0.58
Autoformer	33.24 ± 116.09	37.64 ± 121.51	43.56 ± 137.04	60.79 ± 43.46	64.25 ± 52.02	76.27 ± 80.67	0.77 ± 0.42	1.12 ± 1.02	1.14 ± 0.96
Transformer (Clustering)	28.25 ± 103.07	30.56 ± 114.63	36.98 ± 136.22	51.69 ± 34.32	57.86 ± 43.01	67.71 ± 46.75	0.69 ± 0.35	1.02 ± 0.74	0.99 ± 0.62
Informer (Clustering)	32.66 ± 118.46	35.49 ± 122.8	41.07 ± 134.59	61.22 ± 32.35	63.82 ± 34.71	70.16 ± 46.75	0.77 ± 0.34	1.05 ± 0.78	1.01 ± 0.58
Autoformer (Clustering)	33.51 ± 117.14	36.92 ± 120.62	41.28 ± 134.3	61.23 ± 38.64	64.1 ± 49.24	72.57 ± 75.51	0.77 ± 0.37	1.1 ± 0.98	1.06 ± 0.74
biLSTM	6.53 ± 34.9	8.29 ± 62.98	20.51 ± 98.42	9.83 ± 18.41	14.28 ± 34.5	40.01 ± 34.98	0.15 ± 0.22	0.25 ± 0.54	0.66 ± 0.4
FIGI-Net	**4.34 ± 26.44**	**5.98 ± 61.22**	**19.79 ± 100.08**	**6.98 ± 17.73**	**9.92 ± 37.31**	**39.73 ± 34.99**	**0.1 ± 0.22**	**0.17 ± 0.57**	**0.65 ± 0.42**

granularity and temporal length. Also, COVID-19 data often lacks the seasonality and consistent patterns that Transformer models excel with, making it difficult for them to adapt to sudden changes. Additionally, the shorter data sequences typical in COVID-19 forecasting limited the model's ability to capture complex patterns, while sensitivity to noise in the data further impacted performance. As a result, recurrent neural network-based models proved more effective for COVID-19 forecasting. A detailed description of the performance of each model is provided in Table 1.

Figure 3 (c) focuses on the performance of FIGI-Net against Persistence model. For each time horizon, we generated a violin plot to visualize the RMSE scores of FIGI-Net (in green) 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 model, 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 9.97, in comparison to 79 from Persistence (a 89% reduction).

Despite the high variability in the data, the Wilcoxon signed-rank test comparing paired observations from the same counties under the two different models enabled us to detect significant differences. To further support our statistical comparisons between FIGI-Net and biLSTM (Supplementary Fig. 2 and Supplementary Fig. 5(a)), we conducted supplementary analyses using Cliff's Delta[52] and bootstrap[53] approaches to address variability concerns in RMSE evaluation. Cliff's Delta is a non-parametric effect size measure that quantifies the overlap between two distributions without assuming normality. In parallel, the bootstrap method generates resampled distributions of our error metrics, providing robust estimates of uncertainty and confidence intervals, even with skewed or limited datasets.

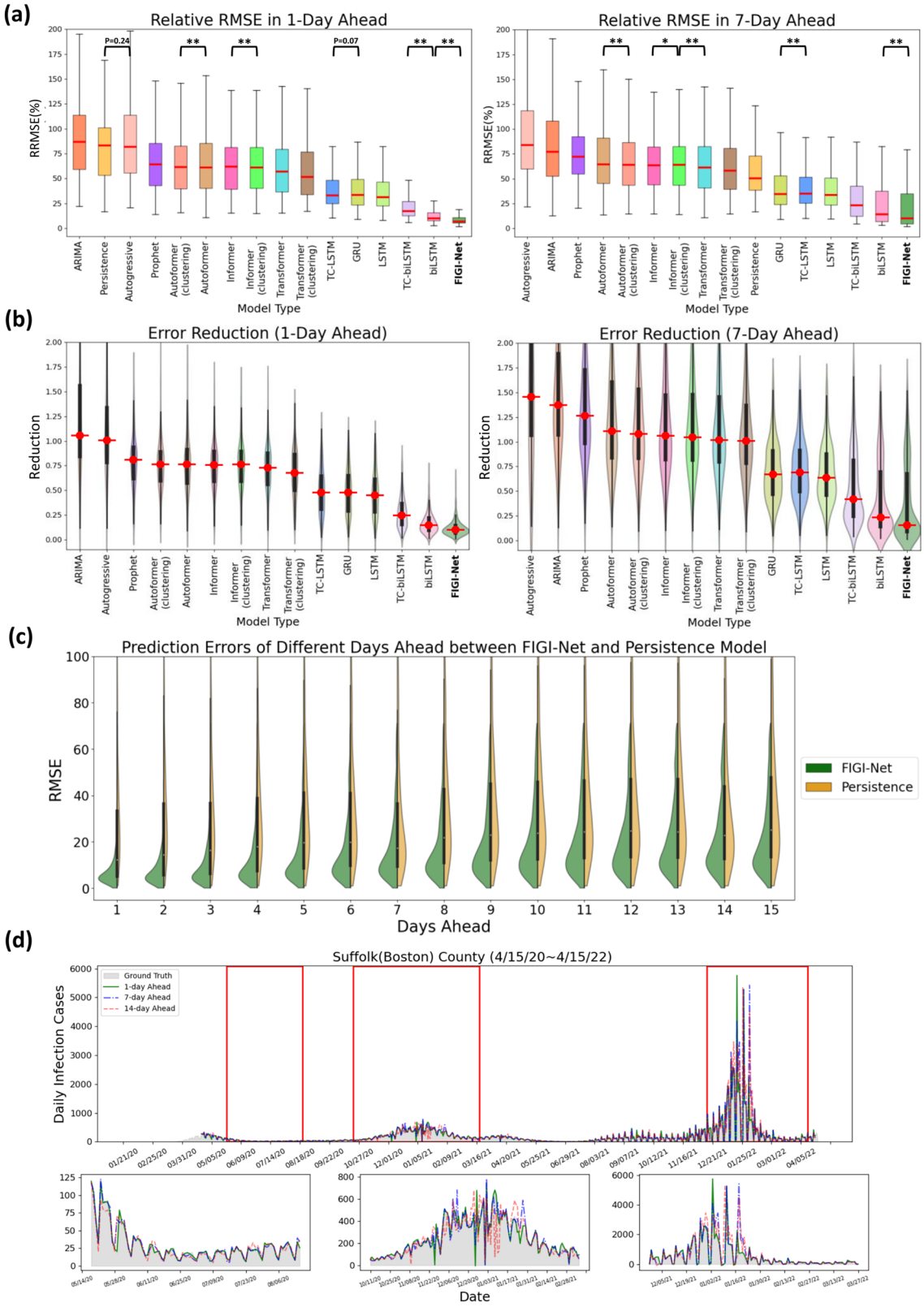
We performed an effect size analysis using Cliff's Delta with 5,000 bootstrapping iterations, as shown in Supplementary Fig. 5(b). The results indicated Cliff's Delta values of 0.224 (95% CI [0.196, 0.251]) for the 1-day forecast and 0.109 (95% CI [0.081, 0.139]) for the 7-day forecast, suggesting small but consistent effect sizes (ranging from 0.1 to 0.3). Notably, the RMSE of biLSTM was larger than that of FIGI-Net. Additionally, to reinforce our two-sided Wilcoxon signed-rank test findings, we performed 5000 bootstrap iterations by subtracting the RMSE of biLSTM from that of FIGI-Net, as shown in Supplementary Fig. 5(c). The results confirmed that the differences between FIGI-Net and biLSTM are statistically significant ($p < 0.01$), demonstrating the reliability of our analysis despite large standard deviations. Collectively, these analyses confirmed the robustness and reliability of FIGI-Net's short-term forecasting performance, showing statistically significant results and small yet meaningful effect sizes.

For a more detailed comparison in short-term forecasting, we also evaluated the proposed model alongside other established models for 1-day and 7-day horizons during the Omicron outbreak. The results, presented in Supplementary Fig. 10 and Supplementary Table 6, demonstrate that FIGI-Net model efficiently handles short-term dynamic changes and achieves approximately a 59% reduction in prediction error. These results underscore the robustness and adaptability of our model in handling rapid and dynamic changes in infection trends, as was characteristic of the Omicron outbreak.

Figure 3 (d) 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, three 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, even when this county exhibited contrasting infection trends 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 the state level

Similar to our county level experiments, 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 high volumes county-level activity as part of its design, we decided to aggregate the county level forecasts into state level (details regarding the performance of FIGI-Net, trained on state-level data, can be found in the Supplementary Table 1). For other deep-learning based models, we adopted the same strategy as FIGI-Net to generate state-level forecasts, as these models require a sufficient amount of data. Additionally, for the rest of the models, we trained them using state-level data to obtain state-level outcomes. Then, we computed the RMSE and RRMSE for each state across 1-day, 7-day, and 14-day ahead horizons, as shown in Fig. 5(a). Our results, also shown in Fig. 4 and summarized in Table 2, show that FIGI-Net was able to score 1149.66 ± 1850.66 and 1935.36 ± 3458.21 in terms of RMSE for the 7-day

Fig. 3 | A summary of the comparative performance of FIGI-Net at the county level. This summary presents a comparative analysis of FIGI-Net's performance at the county level for both the 1-day and 7-day ahead horizon tasks from Oct. 18th, 2020 to Apr. 15th, 2022. **a** Performance of each model over all 3143 counties presented as a box plot. The median is highlighted in red along with the 5th and 95th 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 signed rank test. **b** Error Reduction between each model with Persistence model. Compared to other models, FIGI-Net provides the fewest erroneous forecasting results. **c** 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. **d** shows daily prediction results of our model in 1-day, 7-day and 14-day horizons in Suffolk county in Massachusetts. The example demonstrates our model's ability to provide highly accurate predictions for diverse locations and various time periods. **p*-value < 0.05, ***p*-value < 0.01.

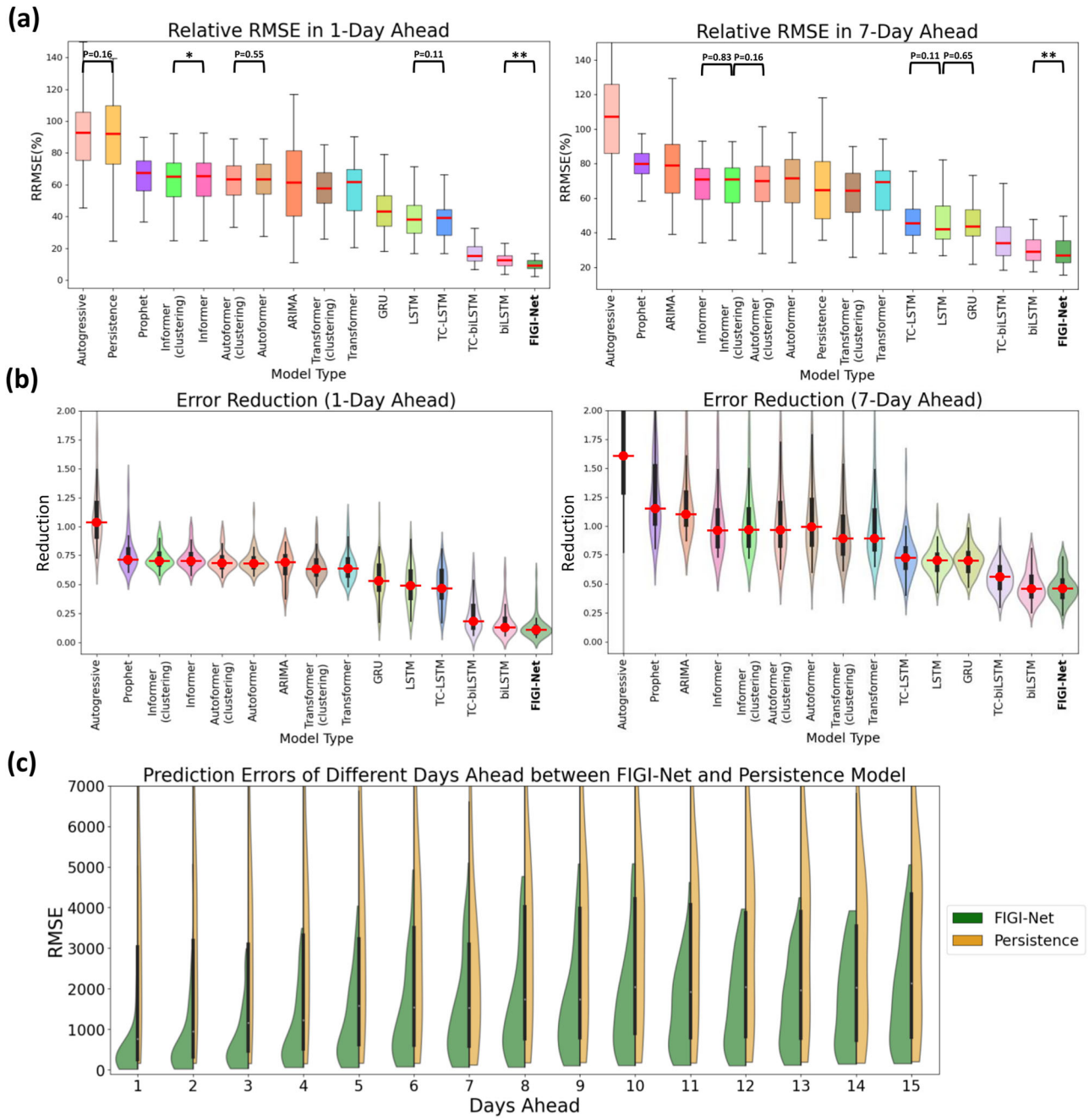


Fig. 4 | A comparative analysis of FIGI-Net's performance at the state level. a At state level, FIGI-Net still exhibits the lowest RRMSE compared to other models for the 1-day and 7-day ahead horizon tasks. **b** Error Reduction at state level displays that FIGI-Net provides lower forecasting errors than other models and reduced errors by at least 53% compared to the Persistence model. **c** Performance comparison of FIGI-Net against Persistence at state level. This comparison also represents that our model consistently displays a much narrower distribution of prediction errors (RMSE) and provides much lower forecasting errors during the first four days. **:*p*-value < 0.01.

and 14-day horizon, correspondingly. On the other hand, the score of the Persistence model was 2571.9 ± 5409.64 and 3330.82 ± 6072.91 (resulting in an error reduction of 54%, and 39% in each case). Next to FIGI-Net, we can see the biLSTM (with a score of 1198.43 ± 1806.38 and 1940.18 ± 3216.91) and TC-biLSTM models (1338.63 ± 2324.23 and 2225.14 ± 4116.09 in RMSE).

During the early stages of the pandemic, we observe the RRMSE values of FIGI-Net were notably higher. For instance, based on Fig. 5(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. Alternatively, the red rectangle in Fig. 5(a) illustrates that the RRMSE errors often increase before the outbreaks or when the infection trend rapidly increases. This observation suggests that the proposed FIGI-Net model may provide early prediction outcomes for outbreaks. Similar trends are also evident in other day-ahead forecasting instances (shown in Supplementary Fig. 6).

National forecasting performance

Figure 5 (b) represents national level COVID-19 official reports contrasted with our model predictions across three different horizons. Similar to the observations in Fig. 3(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%. Table 3 shows the performance between FIGI-Net and other models at national level. Compared to our benchmark models, FIGI-Net improved forecast capacity can lead to a 86.5% 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.

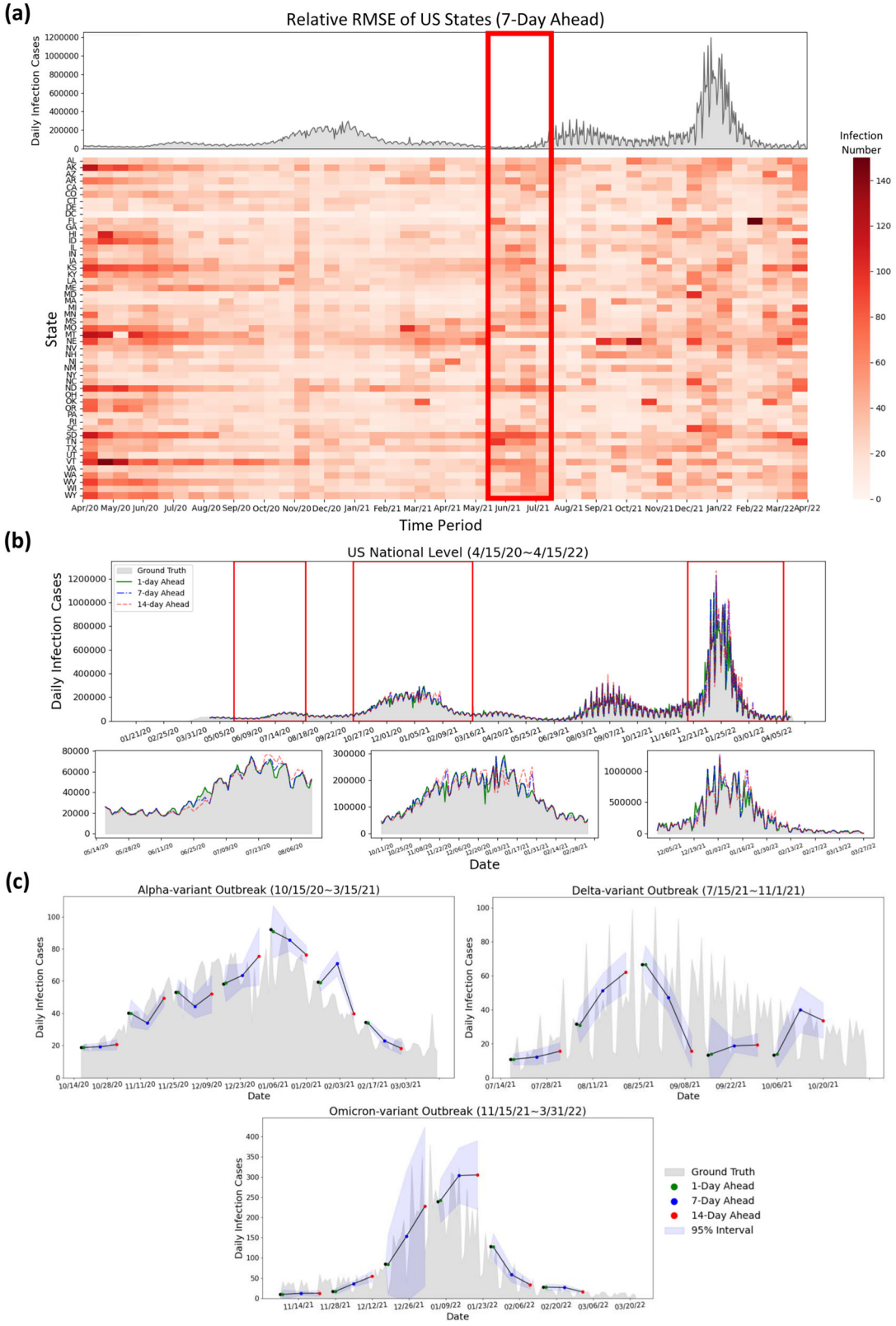
Additionally, we analyzed the daily performance of FIGI-Net at the national level across three main outbreak waves: the Alpha-variant wave (October 15th, 2020 to March 15th, 2021), Delta-variant wave (July 15th, 2021 to November 1st, 2021), and Omicron-variant wave (November 15th, 2021 to March 31st, 2022). By aggregating our county-level forecasts, we depicted the national prediction trends during these waves (Fig. 5(c)). Our model demonstrated an accurate prediction direction, maintaining an average error range of 12.17 infection cases at a confidence level of 95% during the Alpha and Delta-variant waves. During the Omicron-variant wave, the forecasting trend exhibited a wide range of confidence intervals. However, our model still successfully predicted trends up to the 14-day horizon. These above outcomes underscore the robustness of our FIGI-Net model in addressing substantial variations in infection numbers.

Geographical distribution of COVID-19 infection predictions of US counties

We conducted a geo-spatial analysis with the objective to identify possible geographical patterns in the performance of FIGI-Net. Figure 6 illustrates various geographical maps of US counties. The first column of Sections (a), (b), and (c) presents the average number of cases occurring for 3 distinct 1-week periods: the first halves of August 2020, August 2021, and January 2022 (periods just before the Alpha, Delta, and Omicron-variant waves began, respectively). With the objective to demonstrate the forecasting capacity of FIGI-Net, the second column presents the forecast average over the same time periods. Finally, the last column shows the relative RMSE incurred by FIGI-Net. FIGI-Net successfully predicted the COVID-19 activity shown in the 7-day ahead prediction maps compared to the observation maps. However, some counties had larger errors, as observed from the relative RMSE maps (the last column of Fig. 6). 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

Table 2 | Performance metrics with error reduction for the 1-day, 7-day and 14-day ahead tasks at the state level

Model	RMSE			RRMSE(%)			Error reduction		
	1st day	7th day	14th day	1st day	7th day	14th day	1st day	7th day	14th day
Persistence	3563.29 ± 7636.32	2571.9 ± 5409.64	3330.82 ± 6072.91	92.03 ± 29.74	64.74 ± 21.38	79.85 ± 15.76	–	–	–
Autoregressive	3856.01 ± 8189.29	4023.77 ± 1.26x10^{6}	5932.29 ± 3.29 x 10^{10}	92.47 ± 28.45	107.09 ± 1.61 x 10^{4}	160.23 ±: 4.82 x 10^{8}	1.03 ± 0.32	1.62 ± 217.8	2.23 ± 5.72 x 10^{6}
ARIMA	2143.87 ± 5779.11	3260.76 ± 5343.67	3975.23 ± 5926.8	61.36 ± 26.69	78.85 ± 21.21	92.11 ± 14.83	0.69 ± 0.14	1.1 ± 0.4	1.1 ± 0.16
Prophet	2768.8 ± 5128.67	3378.66 ± 5874.52	4121.98 ± 6971.38	67.25 ± 13.31	79.7 ± 9.12	97.91 ± 7.74	0.72 ± 0.23	1.17 ± 0.4	1.21 ± 0.19
GRU	1411.09 ± 3827.44	1460.49 ± 3947.68	1991.35 ± 4503.38	43.29 ± 16.38	43.51 ± 15.36	56.06 ± 14.06	0.53 ± 0.19	0.71 ± 0.18	0.67 ± 0.12
LSTM	1332.83 ± 3342.72	1782.58 ± 3796.64	2028.5 ± 4462.46	38.02 ± 16.33	42.08;± 15.37	53 ± 15.53	0.49 ± 0.19	0.7 ± 0.17	0.67 ± 0.13
TC-LSTM	1311.08 ± 4059.58	1947.45 ± 4131.61	2021.77 ± 4686.08	39.06 ± 16.69	45.5 ± 14.64	52.98 ± 13.81	0.47 ± 0.17	0.73 ± 0.21	0.68 ± 0.11
TC-biLSTM	550.36 ± 1557.85	1338.63 ± 2324.23	2225.14 ± 4116.09	15.36 ± 17.16	33.9 ± 14.38	54.94 ± 15.75	0.18 ± 0.17	0.56 ± 0.15	0.67 ± 0.16
Transformer	2075.91 ± 4175.19	2414.08 ± 4717.82	3252.72 ± 5313.29	61.61 ± 17.85	69.1 ± 16.8	77.88 ± 14.31	0.64 ± 0.16	0.89 ± 0.4	0.93 ± 0.18
Informer	2588.35 ± 5042.39	2757.72 ± 5201.76	3339.13 ± 5617.89	65.42 ± 16.5	70.9 ± 15.32	79.86 ± 10.82	0.7 ± 0.11	0.97 ± 0.39	0.92 ± 0.16
Autoformer	2360.61 ± 4929.34	2686.1 ± 5076.32	3515.96 ± 5702.76	63.41 ± 16.5	71.33 ± 19.74	87.12 ± 17.95	0.68 ± 0.11	1.01 ± 0.43	1.02 ±;0.29
Transformer (Clustering)	2122.01 ± 3858.64	2479.01 ± 4574.4	3269.11 ± 5629.29	57.47 ± 15.7	64.14 ±; 15.12	78.45 ± 11.46	0.63 ± 0.12	0.89;± 0.37	0.92 ± 0.19
Informer (Clustering)	2590.04 ± 5036.38	2749.37 ± 5201.33	3275.72 ± 5613.88	65.07 ± 16.58	70.9 ± 15.16	79.28 ± 10.99	0.7 ± 0.1	0.98 ± 0.39	0.93 ± 0.16
Autoformer (Clustering)	2363.3 ± 4924.15	2714.84 ± 5023.98	3491.41 ± 5451.56	63.45 ± 16.47	69.87 ± 19.37	81.95 ± 17.96	0.68 ± 0.11	0.97 ± 0.4	0.99 ± 0.3
biLSTM	371 ± 1047.57	1189.43 ± 1806.38	1940.18 ± 3216.91	12.44 ± 15.04	28.9 ± 14.01	49.46 ± 17.18	0.13 ± 0.15	0.46 ± 0.18	0.63 ± 0.22
FIGI-Net	**290.31 ± 908.81**	**1149.66 ± 1850.66**	**1935.36 ± 3458.21**	**9.17 ± 13.59**	**26.94 ± 14.48**	**49.12 ± 16.47**	**0.11 ± 0.13**	**0.46 ± 0.15**	**0.61 ± 0.21**



lower RRMSE values in the west coast and northeast regions during these time periods (see, for example, the third column of Fig. 6, while mid-west and south regions of the U.S. tended to display higher errors as the pandemic progressed (refer to Supplementary Movie 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).

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

To further evaluate our approach, we compared the performance of our proposed FIGI-Net model with the COVIDhub_ensemble model[54] (also known as the CDC model). The CDC model employs an ensemble methodology that combines the output of several disease

Fig. 5 | Summary of COVID-19 forecasting results at the state and national levels. **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 predicted infection trends of the proposed model during the Alpha, Delta, and Omicron variant outbreaks with multiple day-ahead predictions. The proposed FIGI-Net model can provide a curve of predicted trends matching the observed report. The range of margin of error became larger when the trend of Omicron-variant outbreak increased.

Table 3 | Performance metrics with error reduction for the 1-day, 7-day and 14-day ahead tasks at the national Level

Model	RMSE			RRMSE(%)			Error Reduction		
	1st day	7th day	14th day	1st day	7th day	14th day	1st day	7th day	14th day
Persistence	94151.72	71901.03	109661.49	51.93	39.65	60.48	—	—	—
Autoregressive	91127.52	1.334×10^8	8.8278×10^{11}	50.26	73574.39	4.869×10^8	0.968	1855.33	8.05×10^6
ARIMA	80821.23	103418.07	149440.99	11.04	29.12	45.42	0.74	1.24	1.18
Prophet	97856.1	144181.6	204519.56	27.38	40.29	55.72	0.9	1.73	1.61
GRU	41298.1	38823.91	57823.73	22.78	21.41	31.89	0.439	0.54	0.527
LSTM	32536.62	39022.06	60314.95	17.94	21.52	33.26	0.346	0.543	0.55
TC-LSTM	37139.2	41291.44	55759.71	20.48	22.77	30.75	0.394	0.574	0.508
TC-biLSTM	16426.44	35465.95	62286.61	9.06	19.56	34.35	0.174	0.493	0.568
Transformer	60811.55	74409.27	115265.9	29.15	35.67	55.26	0.56	0.7	0.66
Informer	76577.56	91711.67	127023.02	36.71	43.97	60.9	0.7	1.1	1
Autoformer	71573.34	85903.96	127543.4	34.31	41.18	61.15	0.66	1.03	1.01
Transformer (Clustering)	62585.83	79163.53	123118.63	30	37.95	59.03	0.57	0.95	0.97
Informer (Clustering)	76380.68	91910.62	126941.02	36.62	44.06	60.86	0.7	1.1	1
Autoformer (Clustering)	71118.17	83600.78	124162.63	34.1	40.08	59.53	0.65	1.01	0.98
biLSTM	13569.52	**27533.97**	51447.97	7.48	**15.18**	28.37	0.144	**0.382**	0.469
FIGI-Net	**12726.38**	28045.28	**50654.95**	**7.02**	15.47	**27.94**	**0.135**	0.39	**0.462**

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. The Persistence model of this task was also included as baseline.

Shown in Fig. 7(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 58.5% in averaged RMSE and 53.28% in averaged RRMSE over the 1 and 2-week ahead horizons, respectively (see Table 4). At the state level (Fig. 7(b)), our proposed model consistently maintained the averaged reduction of 64.55% RMSE value and 64.48% 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.

We also conducted comparative analyses between the COVID-19 forecasts of FIGI-Net with forecasting models officially reported on the CDC website[54] (we selected models, including Microsoft-Deep, JHU_CSSE_DECOM, Karlen-pypm, CovidAnalytics-DELPHI, and MIT_ISOLATMixtures, that provided sufficient infection prediction outcomes for 1-week and 2-week ahead horizons). Following the CDC weekly reporting criteria[55], 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 (Fig. 8(a)). Our findings revealed that most models can accurately predict infection numbers during decreasing trends, but struggle to forecast the correct trend direction and COVID-19 official infection numbers during increasing trends. Interestingly, our FIGI-Net model correctly predicted the increasing direction of the infection trend before the outbreak wave began, from November 2021 to March 2023. Fig. 8b 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.

Comparative analysis of COVID-19 forecasting models during critical time periods

Given that identifying the beginning of a major outbreak is a crucial task in disease forecasting, we assessed 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 20th, 2020 to April 11th, 2022. First, we identified these "critical" time periods as periods where

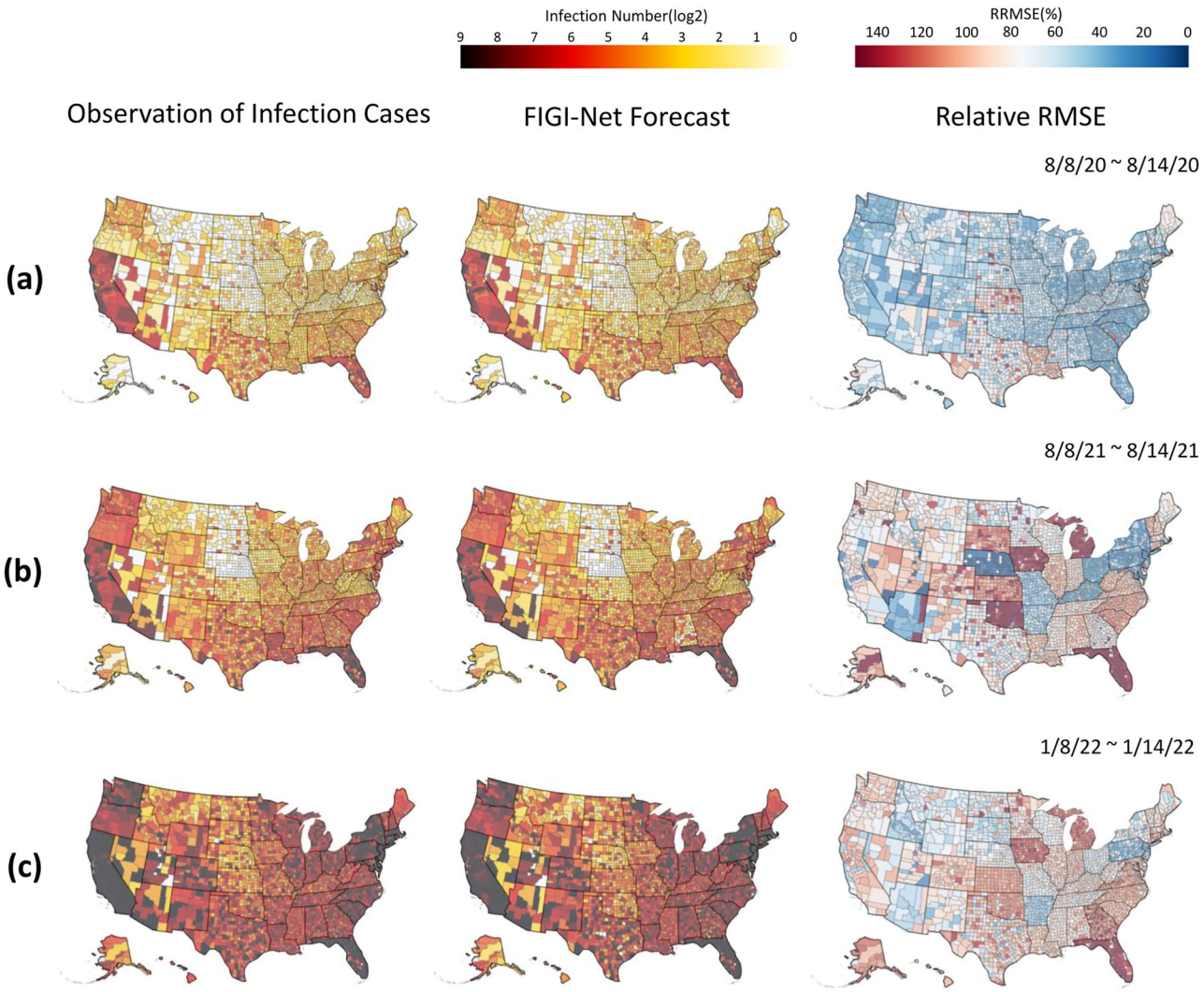


Fig. 6 | Geographical distribution of infection prediction errors of the proposed model during different time periods. The figure shows the infection prediction distribution of the proposed model during different time periods at the county level. The early half of August 2020, August 2021, and January 2022 are represented in (**a**), (**b**), and (**c**), respectively. The first column of (**a**), (**b**), and (**c**) show the observed daily confirmed cases. The second column represent the 7-day (1-week) ahead prediction results, and the corresponding RRMSE score maps are shown in the last column. 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 (**b**) and (**c**) when the status of infections in these two states were severe.

the trend λ of COVID-19 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 (Supplementary Fig. 7 to 9 for more details). Examples of such periods for Massachusetts and New York states are shown in Fig. 8(b). We compared the ground truth and predicted results of the models using slope similarity, RMSE, and MAPE values (see the equation (5) in the section "Model evaluation" 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 (Fig. 8(c)). Table 6 and 7 presented the forecasting performance details among the prediction models and the statistical significance between our model and the others. Comparing with Persistence model, FIGI-Net model improved the RMSE value by at least 43% reduction in 1-week ahead and at least 57% reduction in RMSE in 2-week ahead forecasts. Additionally, our model achieved at least a 45% MAPE reduction in both week horizons and exhibited around 83% 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 have introduced 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 evaluated the forecasting ability of FIGI-Net against state-of-the-art methodologies, including conventional forecasting models (such as ARIMA and Prophet), recurrent neural network architectures (such as GRU, LSTM, TC-LSTM, and biLSTM), and Transformer-based architectures (such as Transformer, Informer, and Autoformer). Our strictly out-of-sample analysis, from October 18th, 2020 to April 15th, 2022, shows that FIGI-Net represents an improvement over existing state-of-the-art models, successfully predicting COVID-19 dynamics at the county, state, and national levels, across multiple time horizons, reaching error reductions of up to 90%, 89.3% and 86.48% in RRMSE, accordingly.

At the county level, FIGI-Net successfully predicted COVID-19 activity, scoring error reductions of up to 90% in comparison to the baseline model, Persistence. FIGI-Net consistently placed as a top 1 performer across the multiple time horizons based on error metrics (RMSE and RRMSE), as presented in Table 1. At the state level, we compared FIGI-Net predictions

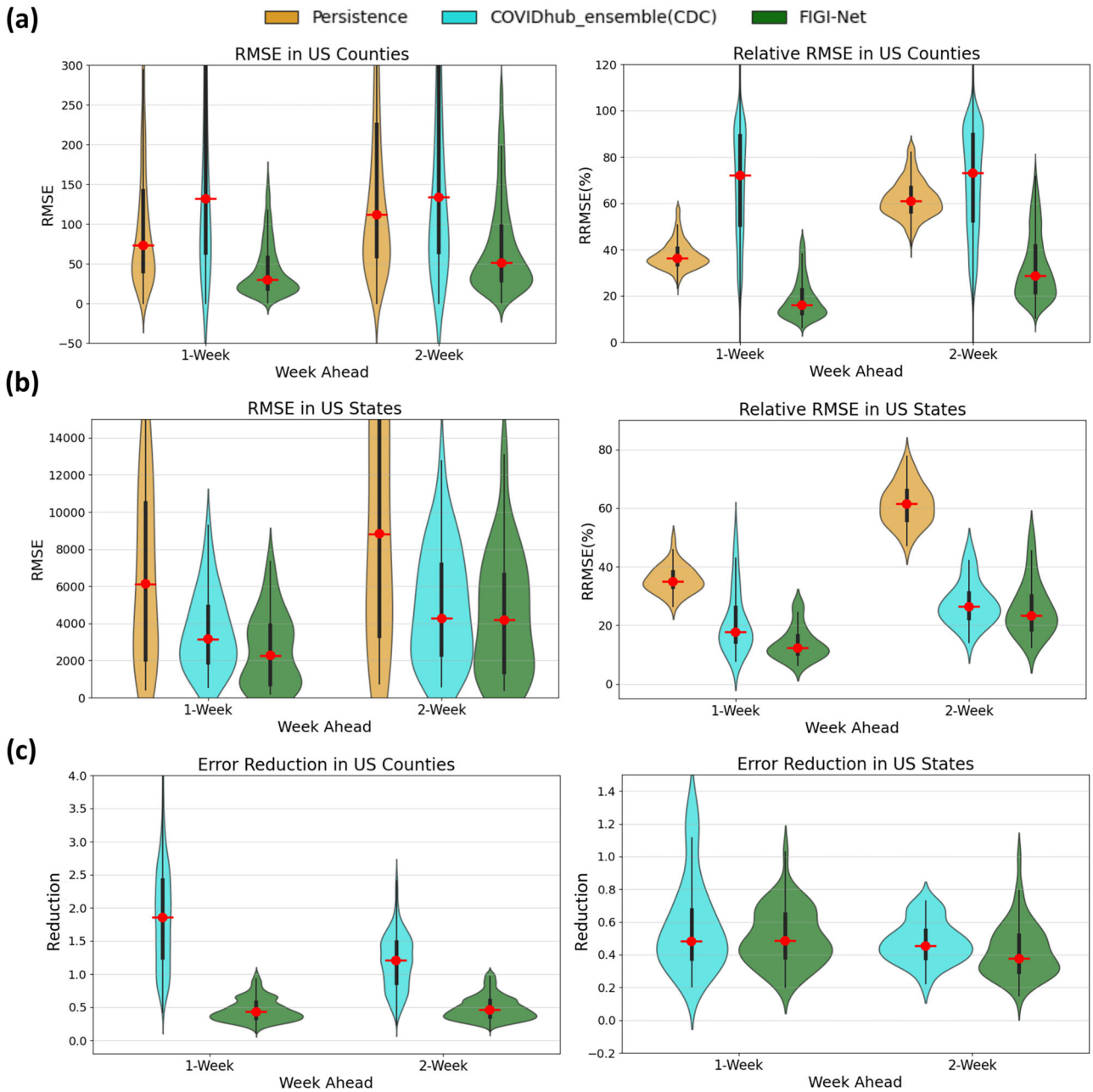


Fig. 7 | Comparison of weekly COVID-19 infection forecasting performance among Persistence model, CDC model, and our proposed model. a The RMSE and relative RMSE values of the three models at the county level. **b** The comparison of prediction errors among these three models at the state level. Our proposed FIGI-Net 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 4.76% averaged reduction in errors observed across various prediction horizons. **c** The error reduction comparison between the CDC model and FIGI-Net. At the county level, FIGI-Net outperforms the CDC model, with errors approximately 58.5% lower than those of the Persistence model. At the state level, FIGI-Net continues to provide a 13% lower error reduction compared to the CDC model.

Table 4 | Comparison of FIGI-Net with CDC and persistence models at county level

Model	1 week				2 week			
	RMSE	RRMSE(%)	Error reduction	*P*-value	RMSE	RRMSE(%)	Error reduction	*P*-value
Persistence	81.24 ± 302.06	37.38 ± 14.77	—	0	125.12 ± 510.7	61.67 ± 11.62	—	0
CDC_ensemble	157.49 ± 1285.95	76.24 ± 78.71	4.63 ± 2.59	0	157.58 ± 1090.97	76.49 ± 66.24	2.83 ± 1.98	0
FIGI-Net	**34.16 ± 190.24**	**17.13 ± 16.49**	**0.36 ± 0.3**	—	**58.43 ± 282.14**	**29.36 ± 21.21**	**0.47 ± 0.3**	—

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

Model	1 week				2 week			
	RMSE	RRMSE(%)	Error Reduction	*P*-value	RMSE	RRMSE(%)	Error Reduction	*P*-value
Persistence	6969.06 ± 11363.25	35.65 ± 11.22	—	8.65×10^{-9}	11019.9 ± 20085.45	62.27 ± 9.51	—	2.1×10^{-9}
CDC_ensemble	3275.07 ± 4823.08	18.93 ± 22.94	0.48 ± 0.23	0.0021	5112.93 ± 8774	26.7 ± 19.18	0.49 ± 0.16	0.0233
FIGI-Net	**3053.46 ± 8315.92**	**12.54 ± 11.29**	**0.34 ± 0.36**	—	**4519.95 ± 10700.57**	**23.57 ± 12.97**	**0.37 ± 0.29**	—

against state-of-the-art models for the 1-day, 7-day and 14-day horizon. Our results showed that FIGI-Net achieved error reductions of 89.3%, 53.76%, and 39.42% in RMSE accordingly, when compared to the Persistence estimate (see Table 2). Finally, at the national level, our model successfully presented an error reduction of up to 86.48% over the Persistence estimates. We attribute the superior performance of FIGI-Net across multiple geographical resolutions 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 subsequently fine-tuned sub-models to mitigate the influence of noise or irrelevant information and increase its own predictive power. This innovative framework better captures the rapidly changing dynamics (see, for example, Fig. 3), ensuring accurate forecasts up to 2 weeks into the future. We also evaluated our model on the top 50 high and low population density counties and compared its performance with Persistence model, Autoregressive model, and recurrent neural network-based models. Our FIGI-Net achieved the two lowest averaged forecasting errors in 1-week ahead predictions for both the top 50 low and high population density counties. These results demonstrate that our model can efficiently handle diverse demographic changes in forecasting, regardless of population density (as shown in Supplementary Fig. 4).

Furthermore, we conducted training on several models exclusively utilizing state-level data and compared the results with those derived from our model, as shown in Supplementary Fig. 1. According to this comparative analysis, the deep learning-based model requires a sufficient amount of effective data size to help reinforce its response to sudden changes in forecasting. Leveraging county-level data provides an ample amount of training information to generate predictive outcomes at a small scale, which can then be aggregated to yield enhanced precision in forecasting at a coarser scale. However, for Transformer-based models, the experimental results indicate that these models are unable to provide accurate predictions in response to sudden changes in forecasting. This may be due to the limited amount of COVID-19 data and the significant variability within the COVID-19 data, which affects the performance of such models. In contrast, our model effectively overcomes these limitations and delivers more accurate outcomes for COVID-19 forecasting.

The analysis of FIGI-Net predictive power against the CDC's official ensemble predictions showed notable improvements seen as substantial reduction in error rates produced by FIGI-Net. At the county level, our model demonstrated more than 50% error reduction, while at the state level, the reduction was at least 63% (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 not only based on an experimental analysis on the predictive power of FIGI-Net as a function of the training window size, but also on our goal of examining how a shorter training duration impacts the model's ability to provide accurate forecasts, particularly during the early stages of the COVID-19 pandemic. Figure 2 illustrates that longer length of training data positively influenced our model's performance, as assessed and identified through the initial six-month dataset for robust evaluation. 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 Fig. 7), we determined the optimal training data length to be 75 days. This time period is short enough to capture changes in disease transmission in a responsive way. In addition, our experiments showed that a training period of at least 60 days produced better 15-day trend predictions, while a 45-day period was effective for 2-week forecasts with shorter training time. Therefore, we chose a 45-day window size to maintain short-term accuracy and improve medium-term predictions.

Upon analyzing the geographical distribution maps presented in Fig. 6, which illustrate our county-level predictions and errors across the U.S., we observed that COVID-19 spread often aligns with state boundaries. This alignment may be attributed to variations in COVID-19 reporting policies and practices, which differ by state government and local jurisdictions[56]. For example, certain states may report cases at different intervals or with varying levels of detail, impacting the consistency and granularity of the data the model receives. Such inconsistencies can lead to variations in the model's predictive accuracy. Additionally, local factors such as public health interventions, community behaviors, and regional policy shifts play a critical role in shaping infection trends within each state[57]. Public health measures like mask mandates, social distancing guidelines, and vaccine rollouts vary widely across states and may change rapidly in response to rising cases, creating dynamic infection patterns that are challenging for the model to capture uniformly. Community behavior, including adherence to public health guidelines and mobility patterns, also introduces variability that can affect model performance.

These local differences 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 Fig. 6(b), 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 Fig. 6(c). 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. Therefore, our model performs optimally in regions with consistent, high-frequency reporting, free from reporting delays, sudden policy shifts, or inconsistencies in data collection. This consistency enables the model to focus on genuine infection trends without interference from extraneous elements that might skew predictions. Additionally, the model's transfer learning approach benefits from homogeneous clusters of regions with similar infection dynamics and public health responses such as comparable socioeconomic factors, healthcare resources, or public health interventions. These similarities enhance forecast accuracy across regions by allowing shared insights.

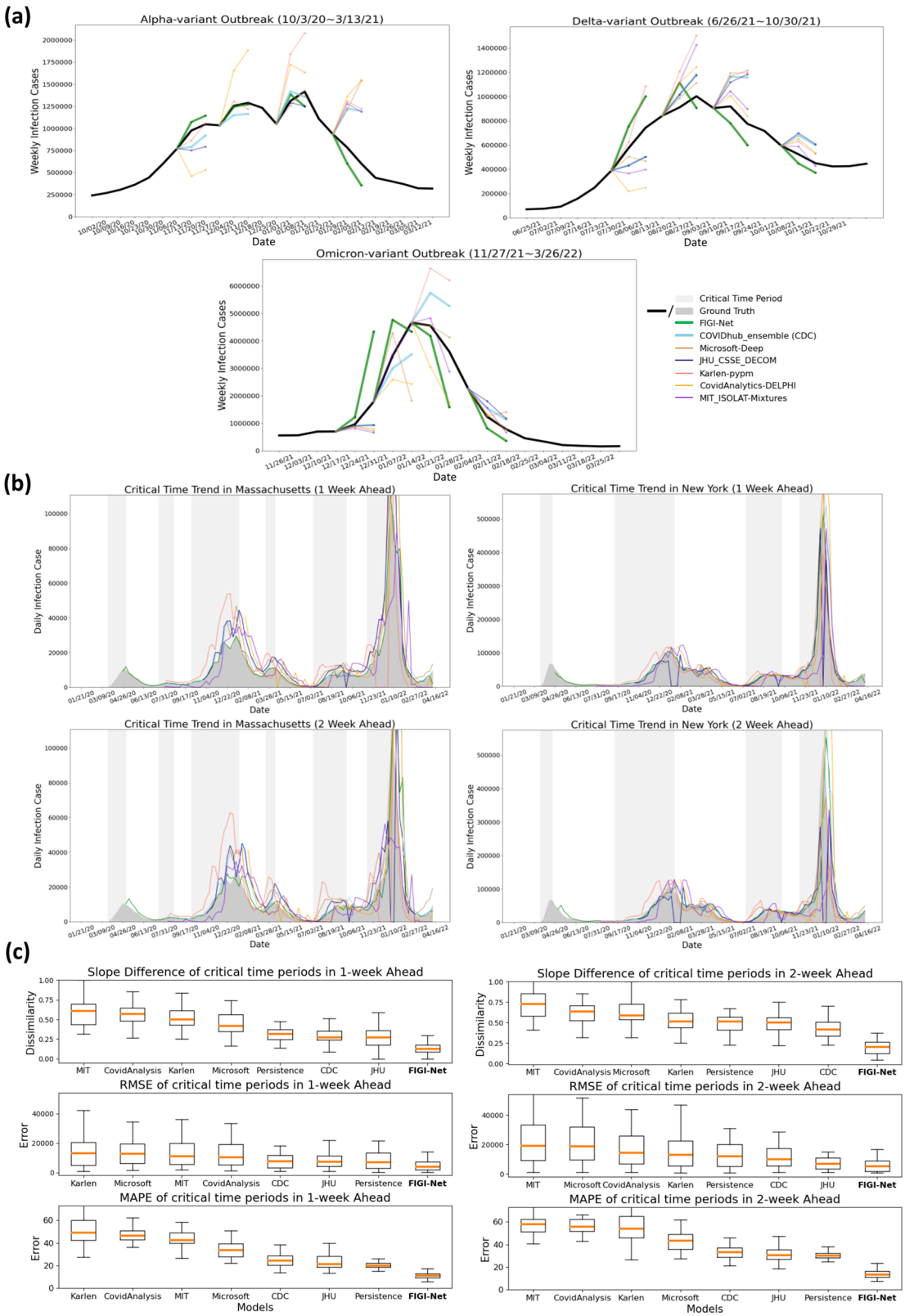


Fig. 8 | Comparison of FIGI-Net model with other state-of-the-art prediction models in predicting weekly and critical time infections. a Comparison of the forecasting results among different forecasting models during three different outbreak time periods at the national level. **b** Examples of COVID-19 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-19 infections, is highlighted in light grey color. **c** Performance evaluation of the forecasting methods during the critical time periods of COVID-19 infection in 1-week and 2-week horizons across the states. Slope difference, 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.

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	0.316 ± 0.079	**0.127 ± 0.076**	0.273 ± 0.108	0.5 ± 0.147	0.57 ± 0.136	0.269 ± 0.134	0.606 ± 0.166	0.421 ± 0.137
RMSE	7015.59 ± 15176.43	**3964.97 ± 6690.04**	7773.23 ± 14058.41	13056.12 ± 14803.5	10394.45 ± 21478.21	7241.44 ± 9418.22	11170.42 ± 22911.14	12775.19 ± 26537.04
MAPE	19.96 ± 3.7	**11.15 ± 2.97**	24.48 ± 27.51	49.08 ± 76.09	46.45 ± 23.76	21.07 ± 57.56	42.64 ± 29.46	33.64 ± 47.7
P-value(v.s. FIGI-Net)	1.99×10^{-9}	-	4.89×10^{-8}	9.31×10^{-10}	5.15×10^{-10}	1.92×10^{-6}	5.45×10^{-13}	5.15×10^{-10}

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	0.316 ± 0.079	**0.205 ± 0.093**	0.419 ± 0.116	0.515 ± 0.147	0.635 ± 0.138	0.501 ± 0.12	0.727 ± 0.182	0.588 ± 0.172
RMSE	12141.1 ± 25299.09	**5126.55 ± 9211.25**	10087.22 ± 22735.07	13032.49 ± 17830.7	14211.7 ± 30659.77	6969.87 ± 10211.86	19078.33 ± 40946.07	18664.36 ± 38065.72
MAPE	30.05 ± 3.98	**13.28 ± 3.9**	33.38 ± 19.77	53.98 ± 78.8	55.88 ± 21.28	30.63 ± 51.35	57.81 ± 24.24	43.47 ± 44.71
P-value(v.s. FIGI-Net)	3.73×10^{-9}	-	7.74×10^{-9}	3.94×10^{-9}	1.57×10^{-9}	0.001	1.32×10^{-11}	1.25×10^{-9}

As depicted in Fig. 8(a) and (b), most assessed forecasting models struggled to predict the direction of future infection trends accurately during multiple time periods, perhaps due to the highly variable transmission rates of the multiple COVID-19 variants. Importantly, FIGI-Net predicted appropriately the trend direction during the initial days of each of the three outbreaks that were studied. 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 dynamic 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 over short-term time periods.

We extended our forecasting methodology to predict daily COVID-19 reported cases in the United Kingdom using multi-horizon forecasting. This involved collecting data from the upper tier local authority of the United Kingdom[58] to train and evaluate our models for internal validation. We compared the forecasting results produced by our models with the methodologies previously introduced in this paper. FIGI-Net consistently provided better forecasting performance in 1-day and 7-day ahead by reducing forecasting errors by approximately 60% when compared to the Persistence model (Supplementary Table 7 to 8, and Supplementary Fig. 11). However, for the 14-day ahead prediction, biLSTM and Transformer performed better, likely due to the much smaller number of counties in the U.K. compared to the U.S., which reduced the training data size in each cluster and limited FIGI-Net's capability for longer-term forecasts. Nonetheless, consistent with our experiments in the U.S., FIGI-Net efficiently handled sudden dynamic changes in short-term forecasting in the U.K. Additionally, to assess generalizability of the model, we conducted cross-dataset experiments using overlapping time periods (the entire year 2021) from both the U.S. and U.K. datasets. Specifically, we performed fine-tuned external validation by pretraining FIGI-Net on the U.S. dataset and fine-tuning it on the U.K. dataset, and vice versa. As detailed in Supplementary Table 9, when the model was pretrained on the U.S. data and then applied to the U.K., MAPE improved but RMSE and RRMSE increased, suggesting partial success in transferring learned patterns while highlighting limitations in handling outliers. Conversely, pretraining on the smaller, less diverse U.K. dataset led to elevated RMSE, RRMSE, and MAPE when validated on the U.S. data, indicating that insufficient data variability can hinder effective knowledge transfer. These findings underscore the importance of dataset size and diversity for robust model performance and emphasize the inherent challenges in applying FIGI-Net across regions with distinct epidemiological and reporting conditions.

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 optimize 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 need for a substantial amount of high-quality training data. Deep learning models rely on large, diverse, and representative datasets to effectively learn underlying patterns and make accurate predictions. Limited or low-quality data can compromise the model's performance. For instance, the smaller sample size and reduced variability in the U.K. dataset constrained the model's ability to capture diverse trends and adapt to sudden outbreaks. Furthermore, variations in data quality, reporting standards, and regional outbreak patterns (e.g., Delta variant surges in the U.S. but not in the U.K., and changes in U.K. reporting frequency around March 2022) affect the model's predictive performance and generalizability. Ensuring the availability of substantial, high-quality training data is therefore crucial for our model's effectiveness, and this could be facilitated by increasing the granularity of infection case data (e.g., building town-level datasets). Addressing these limitations is vital for the model's real-world application, and future research should explore strategies to improve computational efficiency and address regional differences, ultimately enhancing predictive accuracy and generalizability.

In conclusion, the FIGI-Net model represents an improvement in the field of COVID-19 infection forecasting and may serve as a template for future pandemic events. 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 national 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.

Methods

Compliance with ethical regulations

All data used in this research were publicly available, de-identified, and aggregated, with no individual-level or personally identifiable information included. The study did not involve human participants, interventions, or the collection of private information and was therefore exempt from institutional review board (IRB) approval. All analyses were conducted in accordance with applicable ethical standards for the use of publicly available data.

Data collection and cleaning

The data utilized in this study includes the daily COVID-19 cumulative infectious and death cases of U.S. counties, obtained from the Johns Hopkins Center for Systems Science and Engineering (CSSE) Coronavirus Resource Center between January 21st, 2020 and April 16th, 2022[2]. 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 utilized a 7-day averaging method to denoise the daily official COVID-19 cases[59]. We averaged the case number of the current day and the preceding six days to obtain the denoised value. In addition, we extended the experiments by using both the no-averaging method and 14-day averaging method to demonstrate that 7-day averaging method has a better ability to denoise the data while retaining useful temporal information for efficient model training (as shown in Supplementary Table 5). Because CSSE data occasionally had abnormal or missing observations, we selected valid data by including only counties within the continental United States with verified confirmed case data. This resulted in a dataset comprising U.S. 3143 counties between February 2020 to March 2022, which was used as the ground truth for further evaluation. Due to the rapid changes in the COVID-19 situation, we partitioned the dataset into 48 time intervals, each spanning approximately 15 days in length, to train our models separately.

Temporal clustering

Based on the evidence that neighboring COVID-19 dynamics may highly influence local trends (see Noor et al.[60]), our methodology incorporates a two-step spatio-temporal clustering procedure that aims to identify both global and local similarities in COVID-19 activity across U.S. counties. The outcome of this analysis informs our framework by enabling the training of sub-models tailored to the most relevant dynamics for each county. The procedure consists of the following steps: 1) Creating a set of feature vectors representing similarity between each county, 2) Compressing the representation of these features via a

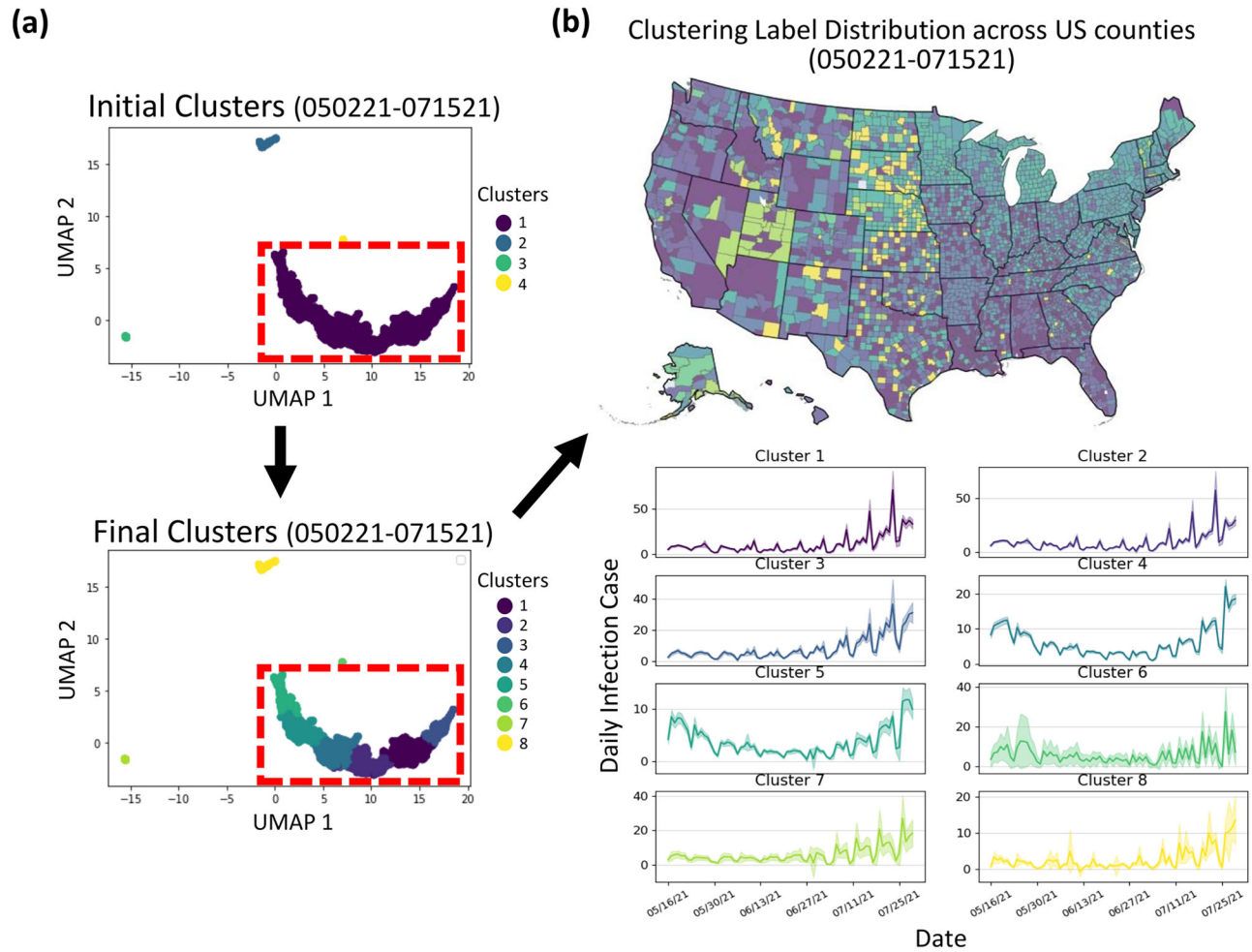


Fig. 9 | An example of temporal clustering for a specific time period. a The UMAP method is used to map temporal correlation features, which are then grouped into initial clusters. The largest cluster (highlighted by the red-dashed rectangle) is further subdivided into subclusters. A total of eight clusters are reordered based on the average infection number over the last 7 days within each cluster. **b** The geographical distribution across U.S. counties demonstrates how temporal clustering captures the relationship between time-series and spatial information. The infection curves of the training data (spanning 75 days) for the clusters, along with a 95% confidence interval, show that this approach provides a collection of highly related yet distinct subclusters, enabling each sub-model to learn efficiently and make accurate predictions.

dimensionality reduction technique (in this case, we apply UMAP), and 3) Performing a two-level clustering process on the resulting reduced dimensions.

Creating the feature vectors. For each time period, we computed the autocorrelation[61,62] of COVID-19 daily confirmed cases and fatality rates, and obtained the cross-correlation between these two data. These two correlation features were concatenated per county, resulting in feature vectors representing the combined correlation information for each county.

Dimensionality reduction. Given that our feature vectors have dimensionality of 301, a more compact representation of these feature vectors is necessary. We transformed our feature vectors using the UMAP method[63], which is a dimensionality reduction technique. We selected UMAP for our dimensionality reduction step due to its ability to preserve both local and global relationships of data during the reduction process. The output of this step is a feature vector of dimensionality of 2.

Two-level clustering. To identify temporal clusters, we used the unsupervised DBSCAN method[64] to extract initial clusters, followed by spectral clustering[65] to identify subclusters within the largest cluster. Figure 9 illustrates a representative example of the process of temporal clustering. The initial clustering captures the global differences in county-level trends, while the subclustering reveals local variations within similar global trends. By doing so, we identified the clusters that include counties from different geographical locations, allowing for the incorporation of relevant local information. Subsequently, the cluster labels was rearranged in descending order of infection risk, based on the average infection count over last 7 days within each cluster. Based on our experiments, we determined the optimal number of clusters to be eight for effective submodel training. When the number of clusters was too small, the submodel tended to learn irrelevant infection features, which adversely affected local prediction accuracy. Conversely, with too many clusters, training data for each submodel became insufficient. Thus, using eight clusters offered a good compromise between providing meaningful local information for accurate predictions and maintaining an adequate volume of training data.

Training sub-models through transfer learning from a global model

Due to the dynamic and rapidly changing trend of COVID-19, the use of deep learning-based time series forecasting models has become essential. The Long Short-Term Memory (LSTM) model, developed from the recurrent neural network, is particularly effective for handling time series

forecasting problems[66]. To address short-term infection variability, we implemented a bidirectional stacked LSTM (biLSTM) model, with the architecture shown in Fig. 1(b). This model leverages the bidirectional processing to capture temporal dependencies in both forward and backward directions, enhancing its ability to manage unexpected, sudden changes. Furthermore, the stacked structure allows our model to better recognize similarities and patterns across the entire historical trend within each time period and across counties.

To achieve accurate and efficient forecasts in each time period, we introduced transfer learning to address the rapid variability and transmission dynamics of COVID-19. We collected 75-day length raw data of all counties prior to each forecast period and applied a 60-day length sliding window, comprising 45 days for training inputs and 15 days for predicted targets, to generate the training dataset for both the entire set of counties and for each temporal cluster. The global model was initially trained on the full dataset to learn generalizable infection patterns. Subsequently, the parameters of this pre-trained global model were transferred to train eight sub-models, each corresponding to a specific temporal cluster. These sub-models were then fine-tuned using county-level data within their respective clusters. During training, the data was randomly split into training (80%) and validation sets (20%). The temporal clustering allows sub-models to optimize the global model's parameters for localized forecasting. 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 was assessed using the RMSE and RRMSE, defined as follows:

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

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

where y, $\hat{y}$, and N represent the observation, the predicted values, and the number of U.S. counties, respectively. To assess the consistency and generalization of the model over time, we use MAPE and error reduction, defined as follows:

$$MAPE = \frac{1}{N}\sum \left|\frac{y-\hat{y}}{y}\right| \times 100. \tag{3}$$

$$Error\ Reduction = \frac{RMSE_{Model}}{RMSE_{Persistence}} \tag{4}$$

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,

$$Slope\ Difference = \frac{2 \times \left|\tan^{-1}\left(Slope\_Observation\right) - \tan^{-1}\left(Slope\_Prediction\right)\right|}{\pi} \tag{5}$$

A lower score indicates that a predicted trend direction closely aligns with the ground truth at a specific time point. We employed linear regression[67] to calculate the slope of each point by using a short time interval that includes the two preceding and two following points. These measurement methods were used to assess the forecasting model's accuracy in capturing rapid trend changes.

For statistical evaluation, the two-sided Wilcoxon signed rank test was employed to assess the statistical significance among the model performances[50]. This test compares paired data from the same counties under two different models, making it suitable for performance metrics without assuming a normal distribution, as shown in Fig. 3(c) and Supplementary Fig. 2. This test does not require the assumption of equal variances between groups, which ensures robustness in cases of large standard deviations and heterogeneous data distributions. By comparing the ranks of observations instead of relying on raw values, the Wilcoxon signed rank test reduces the influence of outliers and variability in data, which ensures fair comparisons between groups. Additionally, the two-sided nature evaluates the statistical significance regardless of the direction of observed differences and provides a reliable p-value. These attributes make the two-sided Wilcoxon signed rank test an appropriate and robust choice for our analysis because it aligns with the characteristics of the data and the objectives of the study. To quantify forecasting variability, the error range function for confidence interval was used and is defined as:

$$margin\ of\ error = Z \times \frac{\sigma}{\sqrt{n}}, \tag{6}$$

where σ and n represent standard deviation of samples and the size of samples, respectively. Z is set to 1.96 for a confidence level of 95%[50].

Data availability

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[2]. 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.

Code availability

All code used to implement the algorithms and conduct the experiments in this study is available online at the following repository: https://github.com/kleelab-bch/FIGI-Net. Users can access the code, along with the instructions for replicating the experiments and obtaining the forecasting results. All code is shared under the MIT License and can be freely accessed and reused for academic and non-commercial purposes.

Received: 9 April 2024; Accepted: 30 March 2025;
Published online: 11 April 2025

References

1. Cucinotta, D. & Vanelli, M. Who declares covid-19 a pandemic. *Acta Biomed.* **91**, 157–160 (2020).
2. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
3. Enserink, M. & Kupferschmidt, K. With covid-19, modeling takes on life and death importance. *Science* **367**, 1414–1415 (2020).
4. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A, Containing Pap. A Math. Phys. Character* **115**, 700–721 (1927).
5. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653 (2000).
6. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science* **368**, 489–493 (2020).
7. Kucharski, A. J. et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *lancet Infect. Dis.* **20**, 553–558 (2020).

8. Lourenco, J. et al. 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* 2020–03 (2020).
9. Dehning, J. et al. Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science* **369**, eabb9789 (2020).
10. Maier, B. F. & Brockmann, D. Effective containment explains subexponential growth in recent confirmed covid-19 cases in china. *Science* **368**, 742–746 (2020).
11. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the covid-19 epidemic? *lancet* **395**, 931–934 (2020).
12. Mitjà, O. et al. Experts' request to the spanish government: move spain towards complete lockdown. *Lancet* **395**, 1193–1194 (2020).
13. Aron, J. L. & Schwartz, I. B. Seasonality and period-doubling bifurcations in an epidemic model. *J. Theor. Biol.* **110**, 665–679 (1984).
14. Harrison, R. Introduction to monte carlo simulation. *AIP Conf. Proc.* **1204**, 17–21 (2010).
15. Chatterjee, K., Chatterjee, K., Kumar, A. & Shankar, S. Healthcare impact of covid-19 epidemic in india: A stochastic mathematical model. *Med. J. Armed Forces India* **76**, 147–155 (2020).
16. Pei, S. & Shaman, J. Initial simulation of sars-cov2 spread and intervention effects in the continental us. *MedRxiv* (2020).
17. Yang, W. et al. Effectiveness of Non-Pharmaceutical Interventions to Contain COVID-19: A Case Study ofthe 2020 Spring Pandemic Wave in New York City. *J. R. Soc. Interface* **18**, 20200822. https://doi.org/10.1098/rsif.2020.0822 (2021).
18. Chinazzi, M. et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**, 395–400 (2020).
19. Lai, S. et al. Effect of non-pharmaceutical interventions to contain covid-19 in china. *Nature* **585**, 410–413 (2020).
20. Cramer, E. Y. et al. The united states covid-19 forecast hub dataset. *Sci. Data* **9**, 462 (2022).
21. Ray, E. L. et al. Comparing trained and untrained probabilistic ensemble forecasts of covid-19 cases and deaths in the united states. *Int. J. Forecast.* **39**, 1366–1383 (2023).
22. Thomas, L. J. et al. Spatial heterogeneity can lead to substantial local variations in covid-19 timing and severity. *Proc. Natl. Acad. Sci.* **117**, 24180–24187 (2020).
23. Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J. & Rosenfeld, R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. *PLoS Computational Biol.* **14**, e1006134 (2018).
24. Ray, E. L. et al. Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRXiv* 2020–08 (2020).
25. Shaman, J. & Karspeck, A. Forecasting seasonal outbreaks of influenza. *Proc. Natl. Acad. Sci.* **109**, 20425–20430 (2012).
26. Sujath, Raa, Chatterjee, J. M. & Hassanien, A. E. A machine learning forecasting model for covid-19 pandemic in india. *Stoch. Environ. Res. Risk Assess.* **34**, 959–972 (2020).
27. Ardabili, S. F. et al. Covid-19 outbreak prediction with machine learning. *Algorithms* **13**, 249 (2020).
28. Hamilton, J. D.*Time series analysis* (Princeton university press, 2020).
29. Hernandez-Matamoros, A., Fujita, H., Hayashi, T. & Perez-Meana, H. Forecasting of covid19 per regions using arima models and polynomial functions. *Appl. Soft Comput.* **96**, 106610 (2020).
30. Lu, F. S. et al. Estimating the cumulative incidence of covid-19 in the united states using influenza surveillance, virologic testing, and mortality data: Four complementary approaches. *PLOS Comput. Biol.* **17**, e1008994 (2021).
31. Lazer, D. et al. Trajectory of COVID-19-Related Behaviors. The COVID States Project, Report. 26, 1–23. www.covidstates.org/reports/trajectory-of-covid-19-related-behaviors (2021).
32. Dutta, S. & Bandyopadhyay, S. K. Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release. *Iberoamerican Journal of Medicine* **2**, 172–177. https://doi.org/10.5281/zenodo.3822623 (2020).
33. Li, D., Huang, G., Zhang, G. & Wang, J. Driving factors of total carbon emissions from the construction industry in jiangsu province, china. *J. Clean. Prod.* **276**, 123179 (2020).
34. Ma, R., Zheng, X., Wang, P., Liu, H. & Zhang, C. The prediction and analysis of covid-19 epidemic trend by combining lstm and markov method. *Sci. Rep.* **11**, 1–14 (2021).
35. ArunKumar, K., Kalaga, D. V., Kumar, C. M. S., Kawaji, M. & Brenza, T. M. 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. *Alex. Eng. J.* **61**, 7585–7603 (2022).
36. Banerjee, S., Dong, M. & Shi, W. Spatial–temporal synchronous graph transformer network (stsgt) for covid-19 forecasting. *Smart Health* **26**, 100348 (2022).
37. Kapoor, A. et al. Examining covid-19 forecasting using spatio-temporal graph neural networks. In *Proceedings of the16th International Workshop on Mining and Learning with Graphs (MLG)*, www.mlgworkshop.org/2020/papers/MLG2020_paper_26.pdf (2020).
38. Pratt, L. Y. Discriminability-based transfer between neural networks. *Adv. Neural Inform. Proces. Syst*. **5**, 204–211 (1992).
39. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks?. *Adv. Neural Inform. Proces. Syst* **2**, 3320–3328 (2014).
40. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724 (2014).
41. Jang, J. et al. A deep learning-based segmentation pipeline for profiling cellular morphodynamics using multiple types of live cell microscopy. *Cell Rep. Methods* **1**, 100169. https://doi.org/10.1016/j.crmeth.2021.100169 (2021).
42. Pan, X. et al. Deep cross-modal feature learning applied to predict acutely decompensated heart failure using in-home collected electrocardiography and transthoracic bioimpedance. *Artif. Intell. Med.* **140**, 102548 (2023).
43. Dekimpe, M. G. & Hanssens, D. M. Empirical generalizations about market evolution and stationarity. *Mark. Sci.* **14**, G109–G121 (1995).
44. Taylor, S. J. & Letham, B. Forecasting at scale. *Am. Statistician* **72**, 37–45 (2018).
45. Ahmed, S. et al. Transformers in time-series analysis: A tutorial. *Circuits, Syst. Signal Process.* **42**, 7433–7466 (2023).
46. Zhou, H. et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 11106–11115 (2021).
47. Wu, H., Xu, J., Wang, J. & Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **34**, 22419–22430 (2021).
48. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat.* **521**, 436–444 (2015).
49. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64**, 107–115 (2021).
50. Niewiadomska-Bugaj, M. & Bartoszynski, R.*Probability and statistical inference* (John Wiley & Sons, 2020).
51. Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M.*Time series analysis: forecasting and control* (John Wiley & Sons, 2015).
52. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* **114**, 494 (1993).
53. Tibshirani, R. J. & Efron, B. An introduction to the bootstrap. *Monogr. Stat. Appl. Probab.* **57**, 1–436 (1993).

54. Cramer, E. Y. et al. The united states covid-19 forecast hub dataset. *medRxiv* https://www.medrxiv.org/content/10.1101/2021.11.04.21265886v1 (2021).
55. Cdc, centers for disease control and prevention, mmwr weeks). https://stacks.cdc.gov/view/cdc/22305. Accessed: 2022-11-15.
56. Stoto, M. A., Woolverton, A., Kraemer, J., Barlow, P. & Clarke, M. Covid-19 data are messy: analytic methods for rigorous impact analyses with imperfect data. *Globalization Health* **18**, 2 (2022).
57. Kaashoek, J. et al. The evolving roles of us political partisanship and social vulnerability in the covid-19 pandemic from february 2020–february 2021. *PLOS Glob. Pub. Health* **2**, e0000557 (2022).
58. Ukhsa data dashboard - covid-19 archive data download. https://ukhsa-dashboard.data.gov.uk/covid-19-archive-data-download (2024). Accessed: 2024-10-01.
59. Lu, F. S., Hattab, M. W., Clemente, C. L., Biggerstaff, M. & Santillana, M. Improved state-level influenza nowcasting in the united states leveraging internet-based data and network approaches. *Nat. Commun.* **10**, 1–10 (2019).
60. Noor, A. U., Maqbool, F., Bhatti, Z. A. & Khan, A. U. Epidemiology of covid-19 pandemic: Recovery and mortality ratio around the globe. *Pak. J. Med. Sci.* **36**, S79 (2020).
61. Wang, C. et al. Deconvolution of subcellular protrusion heterogeneity and the underlying actin regulator dynamics from live cell imaging. *Nat. Commun.* **9**, 1688 (2018).
62. Wang, C., Choi, H. J., Woodbury, L. & Lee, K. Interpretable fine-grained phenotypes of subcellular dynamics via unsupervised deep learning. *Adv. Sci.*, 2403547 (2024).
63. McInnes, L., Healy, J., Saul, N. & Groçberger, L. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
64. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. 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, 226-231 (AAAI Press, 1996).
65. Ng, A., Jordan, M. & Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inform. Proces. Sys.* **14**, 849–856 (2001).
66. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
67. Freedman, D. A.*Statistical models: theory and practice* (Cambridge University Press, 2009).

Acknowledgements

K.L. and M.S. were supported by the National Institutes of Health (NIH) under award number R35GM133725. M.S. was also partially supported by the NIH under award number R01GM130668. M.S. has been funded (in part) by contracts 200-2016-91779 and cooperative agreement CDC-RFA-FT-23-0069 with the Centers for Disease Control and Prevention (CDC). The findings, conclusions, and views expressed are those of the author(s) and do not necessarily represent the official position of the CDC.

Author contributions

K.L. and M.S. conceived and designed the project. T.S. collected the data, designed and implemented the computational framework, and performed the calculation and analysis under the guidance of K.L. and M.S. X.P. and J.J. helped with the initial data and literature collection. T.S. and L.C. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Competing interests

M.S. has received institutional research funds from the Johnson and Johnson foundation, Janssen global public health, and Pfizer. Other authors declare no competing financial or non-financial interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01606-1.

Correspondence and requests for materials should be addressed to Mauricio Santillana or Kwonmoo Lee.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025