EpiGNN: Exploring Spatial Transmission with Graph Neural Network for Regional Epidemic Forecasting

Feng Xie, Zhong Zhang, Liang Li, Bin Zhou(✉), and Yusong Tan

College of Computer, National University of Defense Technology
{xiefeng,zhangzhong,liliang98,binzhou,ystan}@nudt.edu.cn

Abstract. Epidemic forecasting is the key to effective control of epidemic transmission and helps the world mitigate the crisis that threatens public health. To better understand the transmission and evolution of epidemics, we propose EpiGNN, a graph neural network-based model for epidemic forecasting. Specifically, we design a transmission risk encoding module to characterize local and global spatial effects of regions in epidemic processes and incorporate them into the model. Meanwhile, we develop a Region-Aware Graph Learner (RAGL) that takes transmission risk, geographical dependencies, and temporal information into account to better explore spatial-temporal dependencies and makes regions aware of related regions' epidemic situations. The RAGL can also combine with external resources, such as human mobility, to further improve prediction performance. Comprehensive experiments on five real-world epidemic-related datasets (including influenza and COVID-19) demonstrate the effectiveness of our proposed method and show that EpiGNN outperforms state-of-the-art baselines by 9.48% in RMSE.

Keywords: Epidemic Forecasting · Graph Neural Network · Spatial Transmission Modeling · Public Health Informatics.

1 Introduction

Epidemics spread through human-to-human interaction and circulate worldwide, seriously endangering public health. The World Health Organization (WHO) estimates that seasonal influenza annually causes approximately 3–5 million severe cases and 290,000–650,000 deaths.[1] Recently, the coronavirus disease 2019 (COVID-19) has spread over more than 200 countries and territories,[2] causing heavy human losses and economic burdens. Accurate prediction of epidemics is the key to effective control of epidemic transmission and plays an essential role in driving administrative decision-making, timely allocating healthcare resources, and helping with drug research.

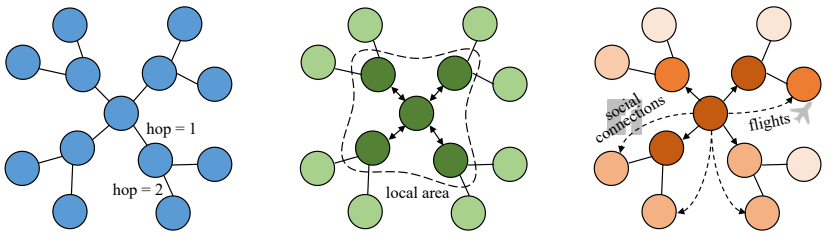
✉ Corresponding author.

[1] https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)

[2] https://covid19.who.int/

A number of studies have investigated epidemic forecasting for decades, aiming to help the world mitigate the crisis that threatens public health. In statistics community, autoregressive (AR) models are widely used in epidemic forecasting [15,3]. In compartment models, the susceptible-infected-recovered (SIR) is the most basic one, many cumulative works in this category are based on its extensions [2,16]. However, the above methods are limited in accuracy and generalization due to their oversimplified or fixed assumptions. Recently, deep learning has achieved tremendous success in many challenging tasks, and various deep learning-based epidemic prediction models [17,1,7] have been proposed, especially models based on emerging graph neural networks (GNNs) [4,11,14]. The core insight behind GNNs is to capture correlations between nodes and model the signal propagation of neighbor nodes. In regional epidemic prediction task, GNN-based approaches model the spread of epidemics by regarding regions as nodes and hidden correlations between regions as edges in a graph structure.

Although both spatial dependencies and temporal information are well exploited, existing methods still face two main challenges. First, the key to GNN-based models is to capture high-quality connections between regions. Using explicit graph structures, such as geographic topology (Fig. 1(a)), does not necessarily reflect the true dependencies or is hard to capture the hidden relationships [19]. Some effective works [11,22] capture potential relationships between regions using specific data (e.g., human mobility) that require struggling with data availability, data accuracy, and data privacy. Due to the excellent feature extraction capability of the attention mechanism [13], several studies [4,7] are mainly dedicated to combining attention mechanism and the latent representation of each region to capture correlations between regions based on similarity. However, owing to the global receptive field of attention mechanism, during aggregating features from other regions, it is prone to causing oversmoothing [9], or bringing noise especially when the data is noisy and sparse in epidemic surveillance [14], which will damage the forecasting performance. Therefore, capturing underlying transmission dependencies between regions reasonably and accurately is crucial to facilitate further improving the prediction performance of GNN-based methods. At the same time, the method we expect should flexibly support both scenarios when rich external information can be collected or not.



(a) Geographic topology. (b) Local transmission effects. (c) Global transmission effects.

Fig. 1. The illustration of geographic topology, local and global spatial transmission effects, where nodes represent regions and edges represent the relationships.

Second, some studies [5,10] have paid much attention to mining the transmission factors of epidemics and assessing the spatial transmission risks of regions, and they suggest transmission risks are meaningful information that provides more practical insights for understanding the spread of epidemics. Spatial transmission risk implies *a potential ability that the epidemic in one region impacts other regions from a spatial perspective*, which is not only important property of regions but also reveals the spatial effects between regions. As shown in Fig. 1, in typical epidemic processes, virus tend to firstly spread in a local range due to intensive mobility of internal elements (e.g., human mobility) between geographically adjacent regions [22] (Fig. 1(b)). Moreover, the epidemic in one region has not only local effects but also spillover effects across regions through complicated social connections [5] (Fig. 1(c)). Thus, modeling spatial effects of regions are beneficial for understanding the spread and evolution of epidemics, which motivates us to investigate how to leverage regional transmission risk to enhance the accuracy and interpretability of epidemic prediction.

To tackle the aforementioned challenges and better understand the spread of epidemics, we propose a novel neural network model, termed EpiGNN, which handles temporal and spatial information through Convolution Neural Network and Graph Convolution Network. In this model, we propose a transmission risk encoding module to characterize spatial effects of regions. Meanwhile, we develop a Region-Aware Graph Learner which takes transmission risk, geographical information, and temporal dependencies into consideration to capture correlations between regions. Our contributions are summarized as follows:

- We design a novel graph neural network-based model for epidemic prediction in which a transmission risk encoding module is proposed that shows how we incorporate local and global spatial effects of regions into the model.
- We introduce a Region-Aware Graph Learner which takes transmission risk, geographical information, and temporal dependencies into account to better explore underlying spatio-temporal correlations between regions.
- We evaluate our model on five epidemic-related datasets. Experimental results show the proposed method achieves state-of-the-art performance and demonstrate the effectiveness of our model. The source code and datasets are available at https://github.com/Xiefeng69/EpiGNN.

The remainder of this paper is organized as follows. We review related works in Section 2. Then we explain the details of our contributions in section 3 and present experiments and results in Section 4. At last, we conclude in Section 5.

2 Related Work

Epidemic forecasting methods. As mentioned above, there has been a large body of work focusing on epidemic forecasting. Essentially, the aim of epidemic forecasting is to predict the number of infection cases for a region at a timestamp based on historical data. In statistics community, autoregressive (AR) models are

widely used in epidemic forecasting [15,3]. In compartment models, susceptible-infected-recovered (SIR) is the most basic one that divides a population into three groups: susceptible, infected, and recovered, and simulates the variations over time between groups. Many cumulative works in this category are based on its extensions [2,16]. Although these methods have a solid mathematical foundation, their accuracy and generalization are limited due to their oversimplified or fixed assumptions, pre-supposed functional form, and careful feature engineering. In recent years, due to its powerful data learning capability, deep learning has been widely adopted in various fields, including epidemic prediction tasks. Wu et al. [17] proposed CNNRNN-Res that firstly applied deep learning for epidemic forecasting. Adhikari et al. [1] adopted deep clustering to help determine the historical season closest to the predicted time point to aid prediction. Jin et al. [6] introduced an inter-series attention-based model to capture similar progression patterns between time series to assist in COVID-19 prediction. Jung et al. [7] designed a self-attention-based approach that cooperates with Long Short-Term Memory (LSTM) for regional influenza prediction.

Graph Neural Network-based models. Graph neural networks (GNNs) have emerged in recent years, such as GCN [8], ST-GCN [20], and demonstrated promising results for extracting the correlation of irregular, non-Euclidean graph data, which make them become powerful tools for understanding the spread and evolution of epidemics. GNN-based epidemic prediction approaches create a graph where nodes correspond to regions of a country, and edge weights correspond to correlations between regions. Deng et al. [4] proposed Cola-GNN that applied an attention mechanism to learn the dependencies between regions based on the latent state of each region learned through Recurrent Neural Networks (RNNs). Panagopoulos et al. [11] took advantage of mobility data across different regions to explore the underlying correlations between regions and adopted message passing neural network (MPNN) combined with LSTM to capture the spatial and temporal evolution of COVID-19. Zhang et al. [22] developed a multi-modal information fusion-powered method that took social connections and demographic information into account to improve COVID-19 forecasting. Wang et al. [14] designed CausalGNN which employed a causal module to provide epidemiological context for guiding the learning of spatial and temporal disease dynamics. Inspired by these works, we aim to explore spatial transmission in typical epidemic processes with GNNs for regional epidemic forecasting.

3 The Proposed Method

3.1 Problem Formulation

We formulate the epidemic prediction problem as a graph-based propagation model. We have a total of N regions (e.g., cities or states). We denote the historical cases data $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_t]$ as training data, where $\mathbf{x}_z \in \mathbb{R}^N$ represents the observed cases value of N regions at time z. Our goal is to predict the future cases value, i.e. $\mathbf{x}_{t+h}$, where h is a fixed horizon with respect to different tasks (e.g., short- or long-term prediction). For every task, we use $[\mathbf{x}_{t-T+1}..., \mathbf{x}_t] \in$

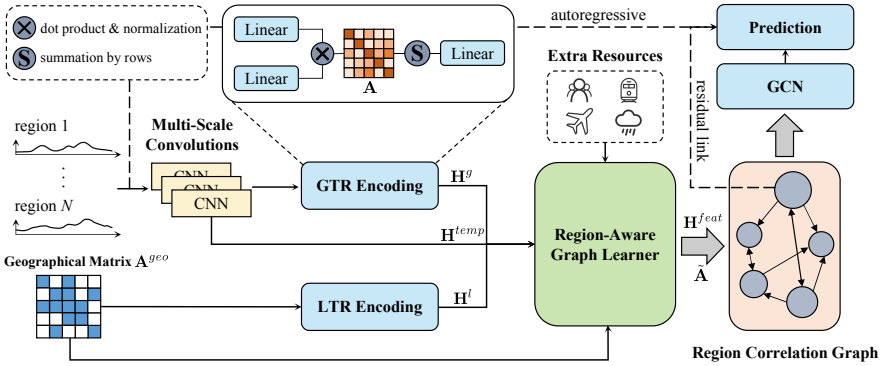


Fig. 2. The overview of our proposed method: EpiGNN.

$\mathbb{R}^{N\times T}$ for a look-back window T to predict $\mathbf{x}_{t+h}$. For a region i, it is associated with a time series $\mathbf{x}_{i:} = [x_{i,t-T+1}, ..., x_{i,t}]$. The proposed method is drawn in Fig. 2. In following sections, we introduce the building blocks for EpiGNN in detail.

3.2 Multi-Scale Convolutions

Convolutional Neural Networks (CNNs) have demonstrated strong feature representation ability and efficient parallel computation in grid data and sequence data that apply learnable filters to capture information behind data. Some works [18,4] suggest that using a set of multi-scale convolutions can capture complex temporal patterns simultaneously. Therefore, in this work, we also adopt multi-scale convolutions with different filter sizes and dilated factors as a feature extractor. We denote convolution filter as $\mathbf{f}_{1\times s,d}$, where s is filter size, d is dilated factor, both s and d are empirically selected. The convolution operation of series $\mathbf{x}_{i:}$ with $\mathbf{f}_{1\times s,d}$ at step j is represented as:

$$\mathbf{x}_{i:} \star \mathbf{f}_{1\times s,d}(j) = \sum_{i=0}^{s-1} \mathbf{f}_{1\times k}(i)\mathbf{x}(j - d \times i), \tag{1}$$

where $\star$ is convolution operator. We use m parallel convolutional layers, each scale with k filters, to generate different feature vectors, and concatenate them after an adaptive pooling layer. We denote $D = (m \times k \times p)$ as the output dimension of multi-scale convolutions, where p is the output dimension of adaptive pooling layer. At last, we obtain the temporal feature $\mathbf{h}_i^{temp} \in \mathbb{R}^D$ for region i.

3.3 Transmission Risk Encoding Module

The epidemic in one region has not only local effects but also spillover effects across regions through complicated social connections [5]. Therefore, We assess local and global transmission risks for regions respectively and encode them as important properties of regions. Essentially, transmission risk encodings indicate spatial structure information which reflects potential spread influence of regions.

Local Transmission Risk (LTR) Encoding. The proximity between regions will lead to a rapid increase in the mobility of internal elements between regions (e.g., human mobility), which will exacerbate local transmission risk. In the geographical network topology, the *degree* is a valuable signal for understanding network structure and describing the centrality of nodes. The more central regions will potentially interact with their surrounding regions more frequently, which leads to significant local spatial effects and is more likely to cause the virus to spread. Hence, we use the degree of each region in geographical topology to measure its local transmission risk. We generate local transmission risk encoding $\mathbf{h}_i^l \in \mathbb{R}^D$ by following equation:

$$\mathbf{h}_i^l = \mathbf{W}^l \cdot d_i + \mathbf{b}^l, \tag{2}$$

where $d_i = \sum_j a_{i,j}^{geo}$ means the degree of region i, and $\mathbf{A}^{geo}$ is the geographical adjacency matrix that indicates the spatial connectivity of regions: $a_{i,j}^{geo} = 1$ means region i and region j are neighbors (by default, $a_{i,i}^{geo} = 1$). $\mathbf{W}^l$ and $\mathbf{b}^l$ are the parameters to transform degree vector $\mathbf{d} \in \mathbb{R}^N$ to encodings.

Global Transmission Risk (GTR) Encoding. Besides geographical adjacent, there are also potential correlations between disjoint regions (e.g., social connections). During the spread of epidemics, it is highly likely that similar progression patterns are shared among related regions because they suffer from the same virus. We believe that if a region has a similar progression pattern to another region, there is probably a dependency between them. Therefore, for global transmission risk assessment, we measure it by the sum of the dynamic correlations based on temporal features of regions, and we call it the *global correlation coefficient* in this paper. Inspired by the self-attention [13], we obtain global correlation coefficients and GTR encodings by following equations:

$$\mathbf{A} = (\mathbf{H}^{temp}\mathbf{W}^q)(\mathbf{H}^{temp}\mathbf{W}^k)^T, \tag{3}$$

$$a_{i,j} = \frac{a_{i,j}}{\max(\|\mathbf{a}_{i:}\|_2, \epsilon)}, \tag{4}$$

$$g_i = \sum_j a_{i,j}, \tag{5}$$

$$\mathbf{h}_i^g = \mathbf{W}^g \cdot g_i + \mathbf{b}^g, \tag{6}$$

where $\mathbf{W}^q$, $\mathbf{W}^k \in \mathbb{R}^{D \times F}$, and $\mathbf{W}^g \in \mathbb{R}^D$ are weight matrices, ϵ is a small value to avoid division by zero. More precisely, we first feed temporal features $\mathbf{H}^{temp}$ to two parallel dense layers and apply a dot product to obtain a correlation distribution matrix $\mathbf{A}$. Then we adopt normalization for each row in $\mathbf{A}$ and calculate global correlation coefficient vector $\mathbf{g} \in \mathbb{R}^N$. At last, we feed g_i to a dense layer to form global transmission risk encoding $\mathbf{h}_i^g \in \mathbb{R}^D$.

3.4 Region-Aware Graph Learner

Capturing correlations between regions by simulating all factors related to the spread of epidemics is troublesome, so we design a Region-Aware Graph Learner (RAGL), which considers both temporal and spatial information to generate a region correlation graph, where nodes correspond to regions, and edge weights correspond to the correlations between regions. We fuse temporal features and transmission risk encodings as nodes' initial attributes $\mathbf{H}^{feat} \in \mathbb{R}^{N \times D}$:

$$\mathbf{h}_i^{feat} = \mathbf{h}_i^{temp} + \mathbf{h}_i^{l} + \mathbf{h}_i^{g}. \tag{7}$$

Existing methods for learning correlations based on attention mechanisms are often symmetric or bidirectional [18]. However, the epidemic transmission is often spread from one region to another, or one region impacts another, so we expect that the learned region correlation graph should not be a completely bidirectional graph. First, we extract dynamic temporal relationships by following equations:

$$\mathbf{M}_1 = \tanh(\mathbf{H}^{temp}\mathbf{W}_1 + \mathbf{b}_1), \quad \mathbf{M}_2 = \tanh(\mathbf{H}^{temp}\mathbf{W}_2 + \mathbf{b}_2), \tag{8}$$

$$\hat{\mathbf{A}} = \mathrm{ReLU}(\tanh(\mathbf{M}_1\mathbf{M}_2^T - \mathbf{M}_2\mathbf{M}_1^T)), \tag{9}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{D \times F}$ are weight matrices. The subtraction term and $\mathrm{ReLU}(\cdot)$ regularize the connectivity of temporal correlation matrix $\hat{\mathbf{A}}$. Next, we capture spatial dependencies utilizing $\mathbf{A}^{geo}$, where we also introduce the *degree* to assess local spatial effects. Specifically, we use the product of the degrees of two adjacent regions as a gate that measures the impacts of local interactions between regions to control spatial dependencies:

$$\mathbf{D}^s = \mathrm{sigmoid}(\mathbf{W}^s \circ \mathbf{d}\mathbf{d}^T), \tag{10}$$

$$\tilde{\mathbf{A}} = \mathbf{D}^s \circ \mathbf{A}^{geo} + \hat{\mathbf{A}}, \tag{11}$$

where $\circ$ is element-wise (Hadamard) product, and $\mathbf{W}^s \in \mathbb{R}^{N \times N}$ is a learnable parameter matrix. The spread of epidemics is associated with many factors (e.g., human mobility, climate). RAGL can flexibly take advantage of external resources that are available to extract dependencies between regions more accurately. We denote external resources as $\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_t]$ where $\mathbf{E}_z \in \mathbb{R}^{N \times N}$ represents external correlation between regions at time step z (e.g., the weight of edge $e_{i,j}^z$ represents the total number of people that moved from region i to region j), we can calculate external correlation matrix by following equation:

$$\mathbf{A}^e = \mathbf{W}^e \circ \sum_{i=0}^{e-1} \mathbf{E}_{t-e}, \tag{12}$$

where e is the look-back window of external resources, and $\mathbf{W}^e$ is a learnable matrix. At last, we sum them up to obtain the region correlation matrix $\tilde{\mathbf{A}}$.

3.5 Graph Convolution Network

Graph Convolution Networks (GCNs) as a kind of GNNs have been proven to be effective methods for learning node representations. In this work, we apply GCN to investigate the epidemic propagation among different regions [8,18,4]. We apply the following equation to update node representations:

$$\mathbf{H}^{(l)} = \sigma(\tilde{\mathbf{D}}^{-1}\tilde{\mathbf{A}}\mathbf{H}^{(l-1)}\mathbf{W}^{(l-1)}), \tag{13}$$

where $\tilde{\mathbf{D}} = \sum_j \tilde{a}_{i,j}$, $\mathbf{W}^{(l)} \in \mathbb{R}^{D\times D}$ is a layer-specific weight matrix, and $\mathbf{H}^{(l)} \in \mathbb{R}^{N\times D}$ is the node representation matrix at l^{th} layer, with $\mathbf{H}^{(0)} = \mathbf{H}^{feat}$. $\sigma(\cdot)$ is the nonlinear function (e.g., exponential linear unit (ELU)).

3.6 Prediction and Objective Function

Due to the nonlinear characteristics of CNNs and GNNs, the scale of neural network outputs is not sensitive to the input. Moreover, the historical infection cases of each region are not purely nonlinear, especially in COVID-19 datasets, showing linear characteristics on the progression patterns of many regions, which cannot be fully handled well by neural networks [21]. To address these drawbacks, some models [3,12] retain the advantages of traditional linear models and neural networks by combining a linear part to design a more robust prediction framework. Therefore, EpiGNN can optionally integrate a traditional AutoRegressive (AR) component as a linear part to obtain the linear result $\hat{\mathbf{y}}^l_{t+h} \in \mathbb{R}^N$:

$$\hat{y}^l_{i,t+h} = \sum_{m=0}^{q-1} \mathbf{W}^{ar}_m x_{i,t-m} + b^{ar}, \tag{14}$$

where q is the look-back window of AR, and $\mathbf{W}^{ar} \in \mathbb{R}^q$ is the parameters in AR component. We concatenate nodes' initial features and the output of the last layer of GCN together, and feed it to a dense layer to obtain the output:

$$\hat{\mathbf{y}}^n_{t+h} = [\mathbf{H}^{(0)}; \mathbf{H}^{(l)}]\mathbf{W}_n + \mathbf{b}_n, \tag{15}$$

where $[;]$ is concatenation operation, and $\mathbf{W}_n \in \mathbb{R}^{2D}$. The final prediction result $\hat{\mathbf{y}}_{t+h} \in \mathbb{R}^N$ of EpiGNN is obtained by summing $\hat{\mathbf{y}}^l_{t+h}$ and $\hat{\mathbf{y}}^n_{t+h}$:

$$\hat{\mathbf{y}}_{t+h} = \hat{\mathbf{y}}^l_{t+h} + \hat{\mathbf{y}}^n_{t+h}. \tag{16}$$

We employ the Mean Squared Error (MSE) to train the model by minimizing the loss. The loss function can be defined as:

$$\mathcal{L}(\theta) = \|\mathbf{y}_{t+h} - \hat{\mathbf{y}}_{t+h}\|_2^2, \tag{17}$$

where $\mathbf{y}_{t+h}$ is the ground truth value, and θ are all learnable parameters in EpiGNN. The pseudocode of the algorithm is described in Algorithm 1.

Algorithm 1 EpiGNN algorithm

Require: Time series data $\{\mathbf{X}, \mathbf{y}\}$ from multiple regions, geographic adjacent matrix $\mathbf{A}^{geo}$, external resources $\mathbf{E}$ (optional).
Ensure: Prediction result $\hat{\mathbf{y}}$.
for each *region* i **do**
  $\mathbf{h}_i^{temp} \leftarrow$ Multi-Scale Convolutions($\mathbf{x}_{i:}$)
  $\mathbf{h}_i^{l} \leftarrow$ Local Transmission Risk Encoding($\mathbf{A}_{i:}^{geo}$)
  $\mathbf{h}_i^{g} \leftarrow$ Global Transmission Risk Encoding($\mathbf{h}_i^{temp}, \mathbf{H}^{temp}$)
end for
for each *region pair* (i, j) **do**
  $\tilde{a_{i,j}} \leftarrow$ Region-Aware Graph Learner($\mathbf{h}_i^{temp}$,$\mathbf{h}_j^{temp}$,$\mathbf{A}^{geo}$,$\mathbf{E}$)
end for
for each *region* i **do**
  $\mathbf{h}_i^{feat} \leftarrow \mathbf{h}_i^{temp} + \mathbf{h}_i^{l} + \mathbf{h}_i^{g}$
  $\mathbf{h}_i^{(l)} \leftarrow$ Graph Convolution Network($\mathbf{h}_i^{feat}, \tilde{\mathbf{A}}$)
  $\hat{y_i} \leftarrow$ Output($\mathbf{x}_{i:}$,$[\mathbf{h}_i^{feat}; \mathbf{h}_i^{(l)}]$)
end for
return $\hat{\mathbf{y}}$

4 Experiments and Analysis

4.1 Experimental settings

Datasets. We conduct experiments on five epidemic-related datasets, three are seasonal influenza datasets and two are COVID-19 datasets. The statistics of datasets are summarized in Table 1. All datasets have been split into training set (50%), validation set (20%), and test set (30%) in chronological order.

Table 1. Statistics of datasets, where SD is standard deviation and granularity means the frequency of epidemic surveillance records.

Datasets	Regions	Length	Min	Max	Mean	SD	Granularity
Japan-Prefectures	47	348	0	26635	655	1711	weekly
US-Regions	10	785	0	16526	1009	1351	weekly
US-States	49	360	0	9716	223	428	weekly
Australia-COVID	8	556	0	9987	539	1532	daily
Spain-COVID	35	122	0	4623	38	269	daily

- **Japan-Prefectures** This dataset is collected from the Infectious Diseases Weekly Report (IDWR) in Japan,[3] which contains weekly influenza-like-illness (ILI) statistics from 47 prefectures from August 2012 to March 2019.
- **US-Regions** This dataset is the ILINet portion of the US-HHS dataset,[4] consisting of weekly influenza activity levels for 10 Health and Human Services (HHS) regions of the U.S. mainland for the period of 2002 to 2017.

[3] https://tinyurl.com/y5dt7stm
[4] https://tinyurl.com/y39tog3h

- **US-States** This dataset is collected from the Center for Disease Control (CDC).[4] It contains the count of patient visits for ILI (positive cases) for each week and each state in the United States from 2010 to 2017. After removing Florida due to missing data, we keep 49 states remaining.
- **Australia-COVID** This dataset is publicly available at JHU-CSSE.[5] We collect daily new COVID-19 confirmed cases ranging from January 27, 2020, to August 4, 2021, in Australia (including 6 states and 2 territories).
- **Spain-COVID** This dataset is collected by [11], consisting of daily COVID-19 cases for 35 administrative NUTS3 regions that were mainly affected by pandemic in Spain from February 20, 2020, to June 20, 2020. We also collect human mobility data in Spain from *Data For Good program*.[6]

Metrics. We adopt Root Mean Squared Error ($RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$) and Pearson's Correlation ($PCC = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}$) as metrics. For RMSE lower value is better, while for PCC higher value is better.

Baselines. We compared the proposed model with the following methods:

- **HA**: the historical average number of cases in observation window T.
- **AR**: the standard autoregression model.
- **LSTM**: the recurrent neural networks (RNN) using LSTM cell.
- **TPA-LSTM** [12]: an attention-based LSTM model.
- **ST-GCN** [20]: a spatial temporal graph neural network.
- **CNNRNN-Res** [17]: a deep learning model that combines CNN, RNN, and residual links for epidemiological prediction.
- **SAIFlu-Net** [7]: A self-attention-based model for influenza forecasting.
- **Cola-GNN** [4]: a deep learning model that combines CNN, RNN and GCN for epidemic prediction.

Implementation Details. All programs are implemented using Python 3.8.5 and PyTorch 1.9.1 with CUDA 11.1 in an Ubuntu server with an Nvidia Tesla K80 GPU. For each task we run 5 times with different random initialization. For all tasks, the batch size is set to 128, the look-back window T is set to 20. The horizon h is set to {3,5,10,15} and {3,7,14} for influenza and COVID-19 prediction respectively in turn. We train the model using Adam optimizer with weight decay 5e-4 and perform early stopping to avoid overfitting. We empirically choose 5 filters: $\{\mathbf{f}_{1\times 3,1}, \mathbf{f}_{1\times 5,1}, \mathbf{f}_{1\times 3,2}, \mathbf{f}_{1\times 5,2}, \mathbf{f}_{1\times T,1}\}$. The range of hidden dimension F is {8,16,24,32}, the number of CNN filters k is searched from {4,8,12,16,32}, the dimension of pooling layer p is chosen in {1,2,3}, the number of GCN layers l is selected from 1 to 5. In COVID-19 task, the model integrates an autoregressive component as a linear part, and the window size q is optimized in {10,20}. In Spain-COVID, we denote EpiGNN$_{exter}$ that utilizes human mobility as external resources, and the look-back window e is searched from {1,2,3}.

[5] https://github.com/CSSEGISandData/COVID-19
[6] https://dataforgood.fb.com/tools/disease-prevention-maps/

Table 2. RMSE and PCC performance of different methods on three datasets with horizon = 3, 5, 10, 15. Bold face indicates the best result of each column and underlined the second-best. $*$ represents that the result is reported in the corresponding reference.

Dataset		Japan-Prefectures				US-Regions				US-States			
		Horizon				Horizon				Horizon			
Methods	Metric	3	5	10	15	3	5	10	15	3	5	10	15
HA	RMSE	2129	2180	2230	2242	2552	2653	2891	2992	360	371	392	403
	PCC	0.607	0.475	0.493	0.534	0.845	0.727	0.514	0.415	0.893	0.848	0.772	0.742
AR	RMSE	1705	2013	2107	2042	757	997	1330	1404	204	251	306	327
	PCC	0.579	0.310	0.238	0.483	0.878	0.792	0.612	0.527	0.909	0.863	0.773	0.723
LSTM	RMSE	1246	1335	1622	1649	688	975	1351	1477	180	213	276	307
	PCC	0.873	0.853	0.681	0.695	0.895	0.812	0.586	0.488	0.922	0.889	0.820	0.771
TPA-LSTM	RMSE	1142	1192	1677	1579	761	950	1388	1321	203	247	236	247
	PCC	0.879	0.868	0.644	0.724	0.847	0.814	0.675	0.627	0.892	0.833	0.849	0.844
ST-GCN	RMSE	1115	1129	1541	1527	807	1038	1290	1286	209	256	289	292
	PCC	0.880	0.872	0.735	**0.773**	0.840	0.741	0.644	0.619	0.778	0.823	0.769	0.774
CNNRNN-Res	RMSE	1550	1942	1865	1862	738	936	1233	1285	239	267	260	250
	PCC	0.673	0.380	0.438	0.467	0.862	0.782	0.552	0.485	0.860	0.822	0.820	0.847
SAIFlu-Net	RMSE	1356	1430	1654	1707	661	870	1157	1215	167	195	236	238
	PCC	0.765	0.654	0.585	0.556	0.885	0.800	0.674	0.564	0.930	0.900	0.853	0.852
Cola-GNN*	RMSE	1051	1117	**1372**	1475	636	855	1134	1203	167	202	241	237
	PCC	0.901	0.890	**0.813**	0.753	0.909	0.835	0.717	0.639	0.933	0.897	0.822	0.856
EpiGNN	RMSE	**996**	**1031**	1441	**1470**	**589**	**774**	**984**	**1061**	**160**	**186**	**220**	**236**
	PCC	**0.904**	**0.908**	0.739	**0.773**	**0.912**	**0.842**	**0.749**	**0.694**	**0.935**	**0.907**	**0.865**	**0.861**

Table 3. RMSE performance of different methods on two COVID-19 datasets with horizon = 3, 7, 14. Bold face indicates the best result of each column and underlined the second-best. - means the forecasting results are not available.

Dataset	Spain-COVID			Australia-COVID		
	Horizon			Horizon		
Methods	3	7	14	3	7	14
HA	167.20	189.90	214.19	2948.48	2777.37	2589.61
AR	165.07	179.51	203.13	85.21	237.73	309.03
LSTM	152.79	177.27	184.44	181.97	315.85	338.34
TPA-LSTM	150.74	183.52	227.95	180.14	220.82	462.78
ST-GCN	162.81	186.21	190.13	253.97	443.01	485.12
CNNRNN-Res	163.75	208.85	219.65	210.23	416.90	488.01
SAIFlu-Net	158.06	200.63	229.62	133.85	277.90	351.14
Cola-GNN	138.34	176.52	203.67	127.59	279.56	326.79
EpiGNN	135.54	162.51	186.41	**71.42**	**153.07**	**287.90**
EpiGNN$_{exter}$	**129.90**	**145.33**	**178.73**	-	-	-

Table 4. Runtime (s) and model size (K) comparison on three influenza datasets when horizon=5. Runtime is the time spent on a single GPU per epoch.

Dataset ($h = 5$)	Japan-Prefectures		US-Regions		US-States	
	Runtime	Params.	Runtime	Params.	Runtime	Params.
ST-GCN	0.18	27K	0.16	26K	0.18	27K
CNNRNN-Res	0.05	13K	0.04	5K	0.06	14K
SAIFlu-Net	0.15	35K	0.10	26K	0.14	32K
Cola-GNN	0.14	9K	0.13	7K	0.15	9K
EpiGNN (ours)	0.10	11K	0.14	9K	0.07	12K

4.2 Prediction Performance

We evaluate each model in short-term (horizon < 10) and long-term (horizon ≥ 10) settings. The experimental results on influenza datasets and COVID-19 datasets are shown in Table 2 and Table 3 respectively. There is an overall trend that the prediction accuracy drops as the prediction horizon increases because the larger the horizon, the harder the problem. The large difference in RMSE across different datasets is due to the scale and variance of the datasets.

We observe that EpiGNN outperforms other models on most tasks. EpiGNN achieves 5.6% and 13.4% lower RMSE than the best baselines in the influenza prediction task and COVID-19 prediction task respectively. In influenza prediction tasks, most deep learning-based models perform better than statistical models (i.e., HA/AR) since they make effort to deal with nonlinear characteristics and complex patterns behind time series. We also notice that statistical model AR is competitive on COVID-19 prediction tasks, especially on Australia-COVID dataset. This could be because of the strong seasonal effects of influenza datasets, which is obviously not the situation in the COVID-19 historical statistics. During COVID-19 period, due to government interventions (e.g., stay-at-home orders, lockdown), the epidemic situations of regions show significant differences. It turns out that a simple linear aggregation over the past case numbers can achieve relatively good performance. EpiGNN also achieves the best performance in COVID-19 datasets attributed to the integration of a linear model. In Spain-COVID, we conduct EpiGNN$_{exter}$ which considers human mobility data as external information in E.q 12 to distill the correlations between regions by providing more practical evidence. The results exhibit that EpiGNN$_{exter}$ is better than EpiGNN, pointing out that external information is helpful for capturing correlations between regions. Table 4 shows the runtimes and number of parameters for each model on influenza datasets. EpiGNN has no obvious adverse effect on training efficiency and well controls the model size to prevent overfitting.

4.3 Ablation Study

- **w/oLTR** stands for EpiGNN without local transmission risk encoding.
- **w/oGTR** represents EpiGNN without global transmission risk encoding.
- **w/oRAGL** indicates EpiGNN using self-attention [13] to capture dependencies between regions instead of Region-Aware Graph Learner (i.e., applying $\tilde{\mathbf{A}} = \text{softmax}((\mathbf{H}^{feat}\mathbf{W}_1)(\mathbf{H}^{feat}\mathbf{W}_2)^T)$).

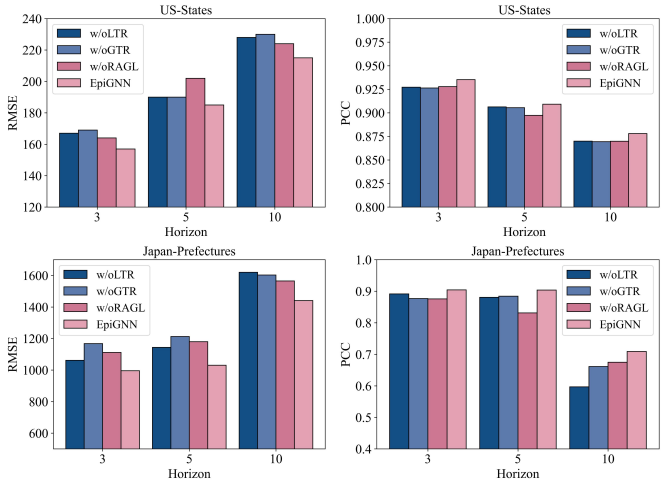


Fig. 3. Results of ablation studies on US-States (top) and Japan-Prefectures (bottom) datasets. For RMSE lower value is better, while for PCC higher value is better.

We perform ablation studies on Japan-Prefectures and US-Regions datasets, and the results measured using RMSE and PCC are shown in Fig. 3. We quantitatively show that the complete EpiGNN can yield the most stable and optimal performance compared to other incomplete models. Compared with using self-attention, RAGL can bring performance gains. The fact can be attributed that RAGL well utilizes spatial and temporal information, which affirms the importance of designing a suitable approach to explore the correlations between regions. In addition, since the captured dependencies are not fully bidirectional, it helps GCN to focus on potentially related regions to overcome the oversmoothing phenomenon [9] and avoid noise accumulation. We also notice that both w/oLTR and w/oGTR cause performance drops, which indicates the positive impacts of transmission risk encodings, and exhibits the effectiveness of modeling transmission risks because they emphasize the spatial effects of regions and provide interpretable evidence on risky areas.

4.4 Parameters Analysis

Number of convolution filters. Different convolution filters learn different features behind data. We evaluate k in range {4,8,12,16,32}, and the results are shown in Fig. 4. Smaller k results in poor prediction performance due to limited representation ability. As k increases, there are more learnable parameters in model and could bring performance gain to a certain extent. We recommend selecting k=12 to achieve a balance between accuracy and computation.

Number of GCN layers. More GCN layers stacked tend to aggregate nodes' features from wider neighborhood ranges. We vary the number of GCN layers

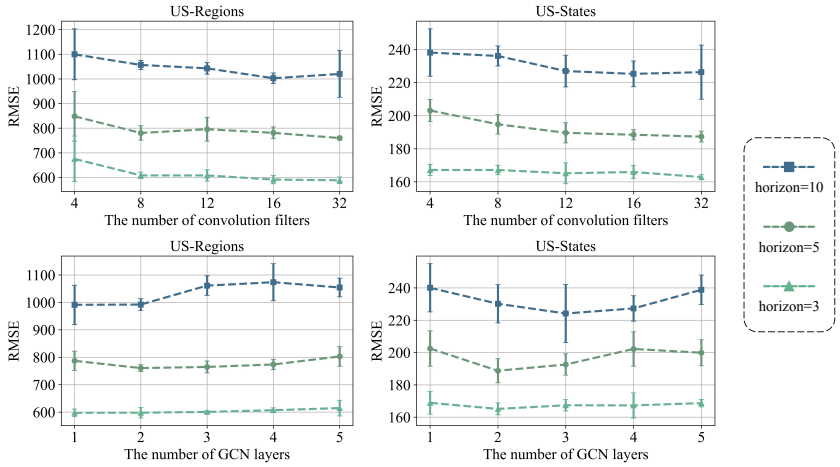


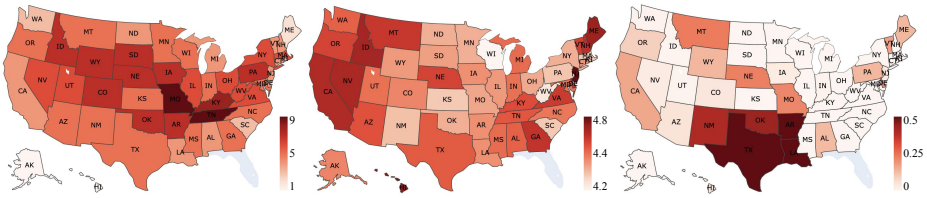
Fig. 4. Parameters analysis results of convolution filter number k (top) and GCN layer number l (bottom) on US-Regions (left) and US-States (right) datasets.

from 1 to 5, and the results are shown in Fig. 4. We observe that smaller l can reach better performance. However, performance drops when l increases reveals that integrating information from irrelevant/weakly-related nodes may result in oversmoothing [9] or bring noises, which will undermine the performance.

4.5 Visualization

We visualize an example with window=(2016/46th-2017/13th) and horizon=5 (week) in US-States dataset, meanwhile, we also provide potential risky regions. Fig. 5(a) is the distribution of degrees in the United States. We notice that the more central/larger region, the greater the degree. Fig. 5(b) is the distribution of global correlation coefficients. Compared with Fig. 5(a), it can be seen that some states (e.g., CA) that are not in the center have high global correlation coefficients. Texas (TX) is the largest and second-most populous state in the U.S. which has a relatively high degree and global correlation coefficient in this case study. We show how Texas is related to other states as drawn in Fig. 5(c). In Fig. 5(c), Texas does not have dependencies with all states. Nevertheless, Texas has relatively significant dependencies with its adjacent regions and also has relationships with some non-adjacent regions.

We visualize the predicted curve of EpiGNN and LSTM in Fig. 6. Compared with LSTM, we observe that EpiGNN fits the ground truth better, and some trends of fluctuation are also predicted better (e.g., WY/DE/VT), while LSTM yields quite inaccurate predictions in some states. We notice that there are similar progression patterns between TX and its adjacent states (e.g., NM/AR/LA), which indicates that local correlations between geographically adjacent regions may be very strong. The correlations drawn in Fig. 5(c) also show that adjacent regions are strongly related, which is consistent with the existing finding [10].



(a) The distribution of degrees. (b) The distribution of global correlation coefficients. (c) The distribution of correlations between Texas and other states.

Fig. 5. Visualization of intermediate results.

5 Conclusions

In this paper, we develop EpiGNN, a novel model for epidemic prediction. In this model, we design a transmission risk encoding module to characterize local and global spatial effects of each region. Meanwhile, we propose a Region-Aware Graph Learner that takes transmission risk, geographical dependencies, and temporal information into account to better explore spatial-temporal dependencies. Experimental results show the effectiveness and efficiency of our method on five epidemic-related datasets. As for future work, we will devote to better predict by considering the time decay effects of spatial transmission.

Acknowledgment. This work is supported by the Key R&D Program of Guangdong Province No.2019B010136003 and the National Natural Science Foundation of China No. 62172428, 61732004, 61732022.

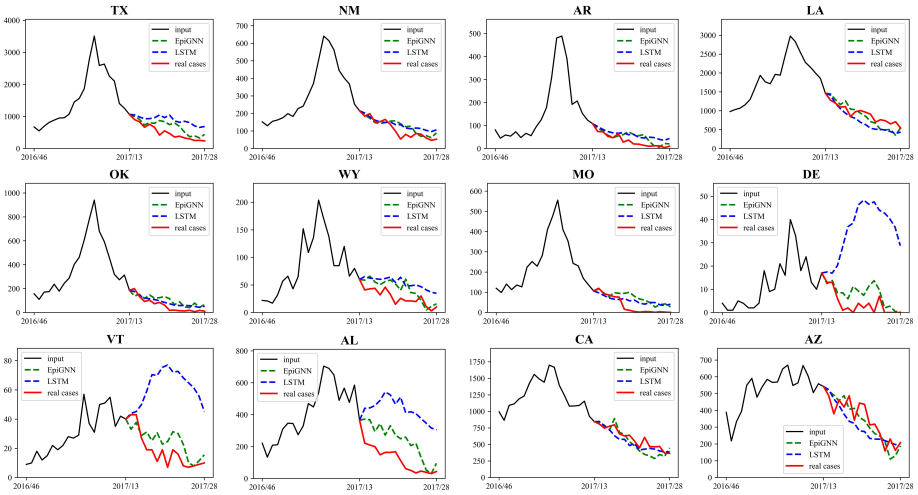


Fig. 6. Predicted curve of EpiGNN (green) and LSTM (blue) for selected states.

References

1. Adhikari, B., Xu, X., Ramakrishnan, N., Prakash, B.A.: Epideep: Exploiting embeddings for epidemic forecasting. In: Proc. of KDD (2019)
2. Aron, J.L., Schwartz, I.B.: Seasonality and period-doubling bifurcations in an epidemic model. Journal of theoretical biology (1984)
3. Chakraborty, T., Chattopadhyay, S., Ghosh, I.: Forecasting dengue epidemics using a hybrid methodology. Physica A: Statistical Mechanics and its Applications (2019)
4. Deng, S., Wang, S., Rangwala, H., Wang, L., Ning, Y.: Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction. In: Proc. of CIKM (2020)
5. Han, X., Xu, Y., Fan, L., Huang, Y., Xu, M., Gao, S.: Quantifying covid-19 importation risk in a dynamic network of domestic cities and international countries. Proceedings of the National Academy of Sciences (2021)
6. Jin, X., Wang, Y.X., Yan, X.: Inter-series attention model for covid-19 forecasting. In: Proc. of SDM (2021)
7. Jung, S., Moon, J., Park, S., Hwang, E.: Self-attention-based deep learning network for regional influenza forecasting. IEEE JBHI (2021)
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
9. Li, Q., Han, Z., Wu, X.M.: Deeper insights into graph convolutional networks for semi-supervised learning. In: Proc. of AAAI (2018)
10. McMahon, T., Chan, A., Havlin, S., Gallos, L.K.: Spatial correlations in geographical spreading of covid-19 in the united states. Scientific Reports (2022)
11. Panagopoulos, G., Nikolentzos, G., Vazirgiannis, M.: Transfer graph neural networks for pandemic forecasting. In: Proc. of AAAI (2021)
12. Shih, S.Y., Sun, F.K., Lee, H.y.: Temporal pattern attention for multivariate time series forecasting. Machine Learning (2019)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Proc. of NeurIPS (2017)
14. Wang, L., Adiga, A., Chen, J., Sadilek, A., Venkatramanan, S., Marathe, M.: Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting (2022)
15. Wang, Z., Chakraborty, P., Mekaru, S.R., Brownstein, J.S., Ye, J., Ramakrishnan, N.: Dynamic poisson autoregression for influenza-like-illness case count prediction. In: Proc. of KDD (2015)
16. Won, M., Marques-Pita, M., Louro, C., Gonçalves-Sá, J.: Early and real-time detection of seasonal influenza onset. PLoS computational biology (2017)
17. Wu, Y., Yang, Y., Nishiura, H., Saitoh, M.: Deep learning for epidemiological predictions. In: Proc. of SIGIR (2018)
18. Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C.: Connecting the dots: Multivariate time series forecasting with graph neural networks. In: Proc. of KDD
19. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling. arXiv preprint arXiv:1906.00121 (2019)
20. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875 (2017)
21. Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. Neurocomputing (2003)
22. Zhang, H., Xu, Y., Liu, L., Lu, X., Lin, X., Yan, Z., Cui, L., Miao, C.: Multimodal information fusion-powered regional covid-19 epidemic forecasting. In: Proc. of BIBM (2021)