Graph Neural Network Modeling of Web Search Activity for Real-time Pandemic Forecasting

Chen Lin
Department of Computer Science
Emory University
Atlanta, USA
chen.lin@emory.edu

Jianghong Zhou
Department of Computer Science
Emory University
Atlanta, USA
jianghong.zhou@emory.edu

Jing Zhang
Department of Computer Science
Emory University
Atlanta, USA
Jing.zhang2@emory.edu

Carl Yang
Department of Computer Science
Emory University
Atlanta, USA
j.carlyang@emory.edu

Eugene Agichtein
Department of Computer Science
Emory University
Atlanta, USA
eugene.agichtein@emory.edu

***Abstract*—The utilization of web search activity for pandemic forecasting has significant implications for managing disease spread and informing policy decisions. However, web search records tend to be noisy and influenced by geographical location, making it difficult to develop large-scale models. While regularized linear models have been effective in predicting the spread of respiratory illnesses like COVID-19, they are limited to specific locations. The lack of incorporation of neighboring areas' data and the inability to transfer models to new locations with limited data has impeded further progress.**

To address these limitations, this study proposes a novel self-supervised message-passing neural network (SMPNN) framework for modeling local and cross-location dynamics in pandemic forecasting. The SMPNN framework utilizes an MPNN module to learn cross-location dependencies through self-supervised learning and improve local predictions with graph-generated features. The framework is designed as an end-to-end solution and is compared with state-of-the-art statistical and deep learning models using COVID-19 data from England and the US.

The results of the study demonstrate that the SMPNN model outperforms other models by achieving up to a 6.9% improvement in prediction accuracy and lower prediction errors during the early stages of disease outbreaks. This approach represents a significant advancement in disease surveillance and forecasting, providing a novel methodology, datasets, and insights that combine web search data and spatial information. The proposed SMPNN framework offers a promising avenue for modeling the spread of pandemics, leveraging both local and cross-location information, and has the potential to inform public health policy decisions.

***Index Terms*—Web search activity, graph neural networks, web-based disease surveillance**

I. Introduction

Over the past decade, there has been an increasing interest in using signals generated from online search activity to predict infectious diseases, such as seasonal influenza and the H1N1 pandemic [1]–[4]. Similarly, since the outbreak of COVID-19, several studies have investigated using online search activity to predict the increase in COVID-19 cases based on the intuition that people with relevant symptoms will search the Web for help [5]–[7]. For example, Fig. 1 shows two time series of COVID-19 related symptom "Rhinitis" search activity and daily confirmed COVID-19 cases in Norfolk, UK during March to May 2020. The peaks of these two curves are highly synchronized and have a strong correlation. In addition, by analyzing the Google search trends [8] and Twitter data, Panuganti *et al.* [6] calculated the relative correlation of online activity concerning different COVID-19 relevant symptoms with the disease incidence and concluded that Google search and tweet frequency regarding "fever" and "shortness of breath" are more robust indicators than "smell loss" for COVID-19 incidence. Meanwhile, Yom-Tov *et al.* [7] analyzed searches for COVID-19 relevant symptoms on Bing search queries from users in England and found that queries for "fever" and "cough" symptoms were the most correlated queries with future COVID-19 cases during the early stages of the pandemic. These studies indicate the feasibility to build COVID-19 forecasting models based on the search activity for COVID-19 relevant symptoms.

Location-specific regression models are the most widely-used method for pandemic forecasting using web search activity. The well-known Google Flu Trends method (the GFT method) applied a linear logit regression model on the aggregated search volume of influenza-relevant queries [1]. Although the GFT method is effective in selecting disease relevant queries, [9] reported that the GFT predictions could be very inaccurate in practice. To overcome this limitation, several studies propose to use linear autoregressive (AR) models with the Elastic Net regularization to learn a sparse model directly on the time series of disease-relevant search queries [4], [5], [10]. For example, Lampos *et al.* [5] have built supervised AR models on COVID-19 relevant search time series and show that they could make predictions preceding the reported confirmed cases and deaths several days ahead. They also show that linear AR models could minimize the concerns that no sufficient data exists at the initial stage of disease outbreaks. Although linear AR models have been built for several respiratory diseases, they have been questioned for lacking the ability of making stable and accurate predictions, mainly because location-specific models tend to be impaired by the irregular change of search activity caused by short-term change in news or media exposure [4].

Furthermore, it is imperative to note that linear logit regression-based methodologies pose challenges in detecting search novelty [11], user interactions [12], [13], and broader geographical connections [14]. As a result, there is a pressing need for research that surpasses location-specific limitations, provides more abundant structure, and possesses the ability to predict and analyze disease tracking effectively.

Graph neural network (GNN) models have been proposed for exploring cross-location dependencies to make more robust prediction for several infectious diseases [15]–[17]. Deng *et al.* [15] propose a graph message neural network with cross-location attention for long-term seasonal influenza prediction with historical disease incidence time series as input and show that the cross-location dependencies in the data improves the model performance. Furthermore, Panagopoulos *et al.* [16] consider the mass mobility data between multiple regions and propose a message passing neural network (MPNN) model to predict the development of COVID-19 based on past disease incidence. In their study, mobility is used as an indicator of spatial connectedness between locations. With MPNN, they update each vertex (region) based on messages received from neighboring regions. According to their results, MPNN has a superior ability to predict the development of diseases compared to multiple baseline models. Although cross-location dependencies in past disease incidence have been explored by prior studies for pandemic forecasting, there is limited work on combining the web search data with geographical graphs for pandemic forecasting, and it remains an open question whether GNNs could outperform location-specific regression models on the web search data.

Additionally, current models based on past disease incidence and mobility data are limited in exploring cross-location dependencies to make more robust predictions for infectious diseases [18]. While Graph Neural Network (GNN) models have been proposed to address this challenge and have shown promising results, there is limited work on combining web search data with geographical graphs for pandemic forecasting [19]. Furthermore, current models have limitations in accurately predicting the development of diseases during the early stages of outbreaks, which is a critical time for taking preventive measures. These limitations have hindered the advancements in pandemic forecasting and surveillance, especially in the context of emerging infectious diseases where timely and accurate predictions are crucial for controlling the spread of the disease.

To address the aforementioned problems, we investigate a novel problem setting, which is to predict the development of disease based on the web search activity using geographical relations between locations. Specifically, we propose a novel self-supervised message passing neural network (SMPNN) framework for pandemic forecasting using the web search activity and the corresponding geographical graph as the input. Our results show that SMPNN extends state-of-the-art GNNs while preserving the advantages of location-specific regression models for pandemic forecasting at the early stage of disease outbreaks. In summary, the main contributions of the paper are:

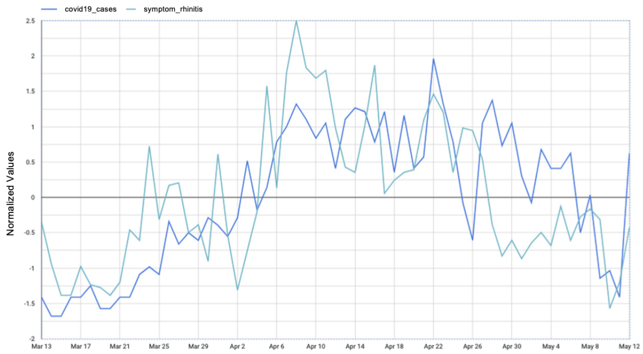


Fig. 1. Two time series from normalized google search volumes of ”Rhinitis” and normalized daily confirmed cases in Norfolk, UK for March to May 2020.

- Identifying a new problem of combining web search activity with a geographical graph for pandemic forecasting.
- Introduction of a novel framework SMPNN, with an MPNN module to learn cross-location features and a location-specific regression module to predict disease incidence.
- Thorough experimental results on two open datasets demonstrating the effectiveness of SMPNN over prior SOTA models.

II. Related Work

This section first outlines previous methods for epidemic forecasting using web search records, showcasing their strengths and limitations. Then, we introduce the Graph Neural Network-based model, which is one of the key models adapted in this paper to overcome the challenges in epidemic forecasting.

A. Web-based epidemic forecasting methods

Google Trends is a widely used web-based epidemic signal for monitoring and predicting outbreaks of infectious diseases. It provides a simple and cost-effective way to track public interest in various topics and keywords related to infectious diseases, thereby offering a unique opportunity to monitor early warning signals of emerging disease outbreaks. Over the past few decades, research on infectious disease prediction using Google Trends has been validated and has shown promising results [1], [20]–[22].

One of the earliest studies to use Google Trends for predicting infectious diseases was published by Ginsberg *et al.* in 2009 [1]. The authors demonstrated that by tracking search volumes for specific keywords such as ”flu” and ”influenza,” Google Trends could accurately follow the temporal and

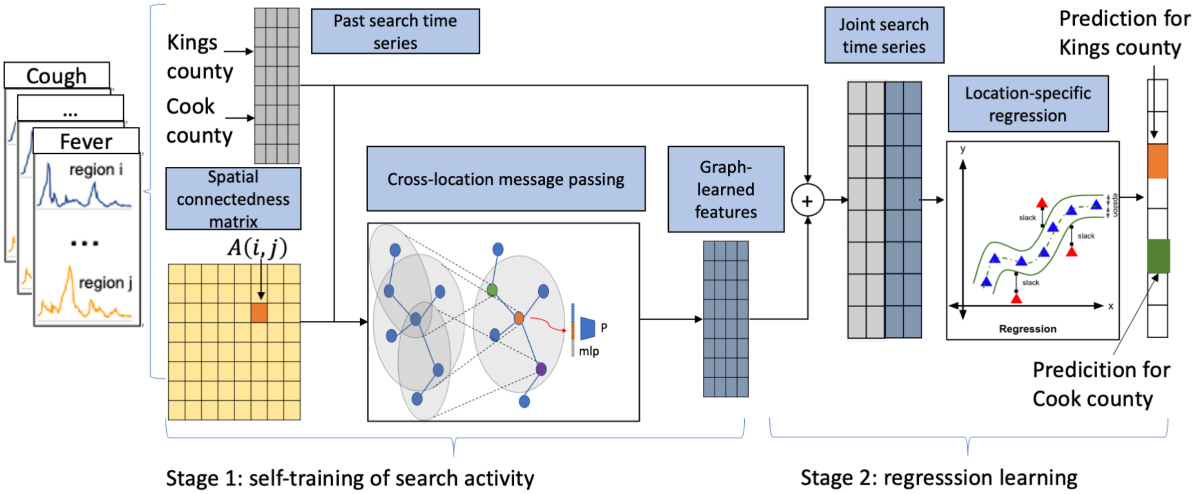

Cough
...
Fever
region i
...
region j
Kings county
Cook county
Past search time series
Spatial connectedness matrix
$A(i,j)$
Cross-location message passing
P
mlp
Graph-learned features
+
Joint search time series
Location-specific regression
Prediction for Kings county
y
x
slack
slack
slack
Regression
Predicition for Cook county
Stage 1: self-training of search activity
Stage 2: regresssion learning

Fig. 2. Illustration of the overall structure of SMPNN model.

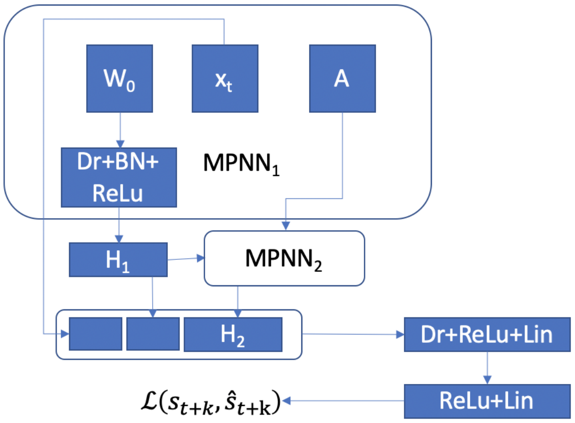


Fig. 3. The illustration of the graph message passing.

geographic patterns of flu-like illnesses. They also found that Google Trends data could predict the onset of flu-like illnesses one to two weeks earlier than traditional surveillance systems, such as the FluView program of the US Centers for Disease Control and Prevention.

Since the release of Google Trends, web-based epidemic forecasting methods have gained significant attention due to the increasing availability of digital data and computational power. Most recent studies on online crowd surveillance for epidemic forecasting have employed various models, with regularized regression models being the most widely used one [4], [5], [10]. Despite their popularity, there have been attempts to explore more sophisticated machine learning models such as random forests and long short-term memory networks (LSTM) [23]. However, these models have limitations in terms of stability and accuracy when predicting disease incidence from web search activity. This is because web search data can be noisy and location-sensitive, and training these models on multiple spatial resolutions may not produce more stable results compared to regularized regression models built for individual locations [23], [24].

Google Flu Trends was a Google project launched in 2008 with the aim of using Google search data to predict the spread of flu. However, after several years of use, it was found to have issues with false positives and over-reporting, as search data may not necessarily reflect actual disease occurrence. With these issues coming to light, Google Flu Trends was shut down in 2015.

The main reason for the decreased accuracy of Google Flu Trends' predictions was that the model and algorithm it used became increasingly different from actual flu outbreaks over time. Specifically, Google Flu Trends used a model based on the volume of keyword searches to predict flu outbreaks. This model used time-series data of keyword searches related to flu, such as ”flu symptoms,” ”cold medicine,” and so on. In the past few years, the algorithm and model of Google Flu Trends were based on previous flu trends. However, as time passed, people's search behavior and habits changed, and the behavior of using keyword searches also changed. This led to data bias in the predictive model, as the correlation with actual flu outbreaks was no longer strong [25].

Therefore, to overcome these limitations, recent research has focused on the use of data fusion techniques to combine multiple sources of data, such as social media, news reports, and healthcare data, to improve epidemic forecasting [26]–[28]. These approaches have shown promise in capturing different aspects of disease spread, such as symptom reporting and healthcare-seeking behavior, and have the potential to improve the accuracy of epidemic forecasting models. However, challenges remain in effectively integrating and modeling these heterogeneous data sources, as they may have different temporal and spatial resolutions and may contain biases and noise.

Furthermore, recent studies in web-based epidemic forecasting also focus on real-time data and adaptive models that can quickly adapt to changing epidemiological conditions [16]. However, this requires the development of robust and scalable methods for collecting, processing, and analyzing large volumes of data in real-time, as well as the ability to quickly update models and predictions based on new data.

The deployment and implementation of epidemic forecasting models in real-world settings requires careful consideration of ethical, legal, and social implications, as well as effective communication and collaboration with public health officials and other stakeholders. While web-based epidemic forecasting methods have limitations, the use of machine learning models, data fusion techniques, and real-time adaptation shows its potential to improve the accuracy and effectiveness of epidemic forecasting.

B. Graph Neural Network-based models

Infectious diseases are an important issue in the global public health field. Over the past few decades, many studies have been devoted to studying epidemic forecasting, where autoregressive models [29] and compartment models such as the susceptible infected-recovered (SIR) models have been widely applied [30]. However, these methods have limited accuracy and generalization due to their oversimplified assumptions. In recent years, deep learning models have achieved great success in various fields and have been widely adopted in epidemic prediction tasks, especially for graph neural network models (GNNs). Using GNNs for infectious disease prediction is a new and emerging research area that involves analyzing disease spread patterns, social networks, and medical data. These studies can not only help healthcare institutions predict disease transmission trends but also provide timely health risk alerts to the public [15], [16], [31], [32].

By design, GNNs utilize graph structures to represent data, which effectively captures the relationships between nodes and enables efficient processing of complex data. In infectious disease prediction, GNNs can establish graph structures of nodes such as cities and populations and model the connections between them to predict the spread of infectious diseases [33].

Recent studies have shown that GNNs can be used to model the spread of infectious diseases. For example, Deng *et al.* proposed using a cross-location attention module in the graph message passing models for long-term influenza-like illness [15]. By modeling the spatial and temporal dependencies in disease transmission, they were able to improve the accuracy of epidemic forecasting. Panagopoulos *et al.* proposed using a message-passing neural network (MPNN) for pandemic forecasting in multiple locations [16]. They found that their model was able to capture the complex relationships between different locations and provide accurate predictions of disease spread. In addition, Xie *et al.* proposed modeling spatial transmission with graph neural networks for pandemic forecasting with local and global encoding modules [32]. By incorporating spatial information into their model, they were able to improve the accuracy of their predictions.

TABLE I
MATHEMATICAL SYMBOLS IN THIS PAPER

Symbol	Remarks
$\mathbb{R}^l$	l-dimensional Euclidean space
$x, \mathbf{x}, \mathbf{X}$	Scalar, Vector, Matrix
G	A geographical graph
V, E	the sets of nodes/edges respectively
$\mathcal{S}$	$\{s_i \mid v_i \in V\}$ the set of nodes attributes
$e_{i,j}$	an edge between v_i and v_j
$w_{i,j}$	the edge weight between v_i and v_j
H	the learned representations
d, l	number of days, number of search terms
Dr	The dropout function
BN	Batch Normalization
ReLu	The ReLU activation function
Lin	The linear dense layer

Although GNNs have shown great potential in capturing complex relationships in data and improving the accuracy of predictions for epidemic forecasting, training GNNs requires a large amount of labeled data, which is a key bottleneck to achieving better predictive performance. Unsupervised and semi-supervised self-learning methods have been increasingly attracting attention in deep learning, as they do not require much labeled data. Unlike supervised learning, unsupervised learning extracts information from unlabeled data, which can greatly reduce the workload of manual labeling. Self-supervised learning is an unsupervised learning method in which the model learns information from the data itself, rather than relying on annotated data, which has been applied in the training of GNNs in recent studies [34], [35].

Using self-supervised learning methods has another advantage in that it can handle missing and noisy epidemic data. These issues are common in real-world epidemic data, such as low data quality or missing data due to privacy concerns. Using self-supervised learning methods can alleviate these issues and improve prediction accuracy. Therefore, it is promising to apply self-supervised learning methods to epidemic forecasting tasks in graph neural networks. These methods could help us better understand the process of epidemic transmission and evolution, and improve prediction accuracy and generalization ability.

III. METHODOLOGY

In this section, we first formulate the problem. Then we present the proposed neural network architecture and how it aggregates features for predicting the development of COVID-19. The important notations are summarized in Table I.

A. Problem Formulation

Input. We construct the daily snapshot of the search activity network as a (geographical) Graph $G = (V, E)$, where $n = |V|$ denotes the number of nodes and the weight $w(u, v)$ of the edge (u, v) represents spatial connectedness index between vertex u and vertex v. Specifically, for a given country, the nodes represent its subregions, and the edge weights are calculated by the mobility and social connectedness between the nearby sub-regions.

Spatial Aggregation. Given that people in nearby regions could move and contact with each other, the search activity in one region could be influenced by nearby regions. Therefore, the spatial connectedness index between the regions u and v at time t could be multiplied by the search activity $s_u^{(t)}$ of region u at time t to generate a relative value which represents the extent to which search activity in region v is influenced by region u at time t. Specifically, let $\mathbf{x}_u^{(t)} = \left(s_u^{(t-d)}, \ldots, s_u^{(t)}\right)^\top \in \mathbb{R}^{d*l}$, where $s_u^{(t)} \in \mathbb{R}^l$ is a vector of node features, which consists of the normalized search volume of l search terms of the past d days in region u. We use the search volumes of multiple days rather than considering only the previous day for prediction because search volumes vary greatly between days. In summary, the spatial aggregation process could compute a feature vector for each region with the following formula:

$$\mathbf{A}\mathbf{X}^{(t)} = \begin{bmatrix} w_{1,1}^{(t)} & w_{2,1}^{(t)} & \cdots & w_{n,1}^{(t)} \\ w_{1,2}^{(t)} & w_{2,2}^{(t)} & \cdots & w_{n,2}^{(t)} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1,n}^{(t)} & w_{2,n}^{(t)} & \cdots & w_{n,n}^{(t)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^{(t)} \\ \mathbf{x}_2^{(t)} \\ \vdots \\ \mathbf{x}_n^{(t)} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_n \end{bmatrix} \tag{1}$$

where $\mathbf{A}$ is the spatial connectedness matrix of $G^{(t)}$ and $\mathbf{X}^{(t)}$ is a matrix whose rows consists of the node features of each region. After spatial aggregation, $\mathbf{z}_u \in \mathbb{R}^{d*l}$ is a vector that aggregates the search activity within and towards region u.

Output. The goal of our work is to predict y_u^{t+k}, which is the reported number of COVID-19 cases for region u at k days after day t.

B. Model Designs

The main aim of our work is to model people's web search activity in graph G from real-time data, and measure the deviations from their search behavior to facilitate disease surveillance. To meet this goal, we design a two-stage framework as shown in Fig. 2: (1) Self-supervised MPNN module to generate cross-location features. and (2) Location-specific regression module for disease prediction based on past search volumes and graph-generated search features.

1) Self-supervised MPNN module: MPNN framework represents a family of graph neural network models which use the message, update and readout functions to learn representation from the nodes in the graph [36]. As shown in Fig. 3, we apply two neighborhood aggregation layers in the network and each layer learns from the graph structure and the node representation from the previous layer. We calculate the node representation for each layer using the following formula [31]:

$$\mathbf{H}^{i+1} = f\left(\tilde{\mathbf{A}}\mathbf{H}^i\mathbf{W}^{i+1}\right), \tag{2}$$

where $\mathbf{H}^i$ denotes the node representation matrix of the previous layer. $\mathbf{H}^0 = \mathbf{X}$, represents the initial feature matrix. $\mathbf{W}^i$ denotes the parameter matrix of layer i and f is ReLU activation function. We train the parameter matrix using following loss function:

TABLE II
DATASET STATISTICS FOR ENGLAND AND USA.

COUNTRY	TIME	AVG CASE	MAX CASE	SD
ENGLAND	3/20-5/20	25.04	152.58	20.17
USA	9/20-12/21	279.56	10682.70	477.91

$$\mathcal{L} = \frac{1}{n}\sum_{u \in V}\left(s_u^{(t+k)} - \hat{s}_u^{(t+k)}\right)^2, \tag{3}$$

where s_u^{t+k} denotes the search volume of the search terms for region u at day $t+k$ and $\hat{s}_u^{(t+k)}$ denotes the predicted search volume of the search terms at day $t+k$.

2) Location-specific regression module: At the second stage, we apply location-specific regression models $f(\cdot)$ to predict disease incidence based on past search volumes and graph-generated search features for L symptoms. We predict the disease incidence according to the formula as shown below:

$$\hat{\mathbf{y}}_{\mathbf{u}} = f(\mathbf{S}_{\mathbf{u}}, \beta_u) + \epsilon_u. \tag{4}$$

When we use linear autoregressive model as the regression module, we optimize the model according to the following formula:

$$\arg\min_{\mathbf{w}_{\mathbf{u}}, b_u}\left(\|\mathbf{y}_{\mathbf{u}} - \mathbf{S}_{\mathbf{u}}\mathbf{w}_{\mathbf{u}} - b_u\|_2^2\right), \tag{5}$$

where $\mathbf{y}_{\mathbf{u}} \in \mathbb{R}$ is the reported cases in region u, and $\mathbf{w}_{\mathbf{u}} \in \mathbb{R}^{2*l*d}, b_u \in \mathbb{R}$ denote the feature weights and regression intercept, respectively. Note that the time index of $\mathbf{S}_{\mathbf{u}}$ is omitted for the simplicity of notation. In fact, for a specific search term out of l search terms, we use the search volumes of past d days and the graph-generated features from MPNN module of past d days.

3) End-to-end training pipeline: As mentioned above, we have two options for training the SMPNN model. The first way is to train SMPNN algorithm in an end-to-end way, where we apply a regression layer on top of the MPNN module. The pseudocode of the end-to-end SMPNN algorithm is described in Algorithm 1. The second way is to train the self-supervised MPNN module and location-specific regression module separately. Compared to end-to-end training of SMPNN, the second option preserves more location-specific information and has the flexibility to choose different regression models for location-specific regression.

IV. EXPERIMENTS

A. Dataset

In this subsection, we introduce how we build our datasets from England and the US. Specifically, we collect England data from an open benchmark dataset provided by Panagopoulos *et al.* [16] and collect the US data from the Google COVID-19 open dataset [8]. Their data statistics are shown in Table II. All disease cases are normalized as cases per million people.

TABLE III

MEAN ABSOLUTE ERROR FOR COVID-19 FORECASTING IN NUMBER OF CASES PER MILLION PEOPLE PER REGION.

Model		**England**							**USA**						
Days ahead (k)		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ARMA		17.45	16.81	17.17	17.46	17.60	16.55	15.77	233.1	235.2	237.2	237.5	235.2	228.8	226.8
RF		13.07	14.03	14.79	14.35	13.99	13.34	13.03	245.5	243.9	247.0	242.8	245.2	239.4	234.9
SVR		15.11	14.21	**14.07**	14.00	14.23	13.95	13.60	215.0	**219.7**	223.0	220.3	217.1	209.1	207.4
MPNN		18.19	17.52	18.07	18.71	18.35	18.03	18.96	221.7	221.3	221.6	215.8	226.0	217.8	222.2
end-to-end SMPNN		19.20	19.00	19.28	19.95	18.71	19.28	20.33	221.4	226.9	226.6	227.3	213.9	221.7	219.7
SMPNN	w/ ARMA	16.26	16.50	16.53	17.18	16.78	15.76	16.34	228.7	239.5	232.6	227.6	224.7	216.7	216.7
	w/ RF	**12.83**	14.02	14.62	**13.99**	**13.36***	**12.67***	**12.81**	250.6	244.5	241.4	238.1	237.6	233.4	233.6
	w/ SVR	14.40	**13.85**	14.25	14.13	14.07	13.57	13.45	**212.1**	228.7	**216.4**	**207.0***	**202.2***	**198.4***	**197.2***
Relative Improvement		↑1.8%	↑1.3%	↓1.3%	↑0.1%	↑4.5%	↑5.0%	↑1.7%	↑1.4%	↓4.1%	↑2.4%	↑4.1%	↑6.9%	↑5.1%	↑4.9%

Notes: The numbers are computed as the average of 21 runs/days for the UK and 11 runs/months for the USA, where $^*p < .05$

Algorithm 1 SMPNN algorithm

```
Require: Time series data ${X, y}$ from multiple regions, spatial connectedness matrix $A$
Ensure: Model parameters $\Theta$, prediction result y
for each epoch do
    Randomly sample a mini-batch
    for each region i do                    ▷ Self-supervised process
        $\mathbf{h}_i \leftarrow$ Graph Message Passing $(\mathbf{x}_{i:}, A)$
        $\hat{s}_i \leftarrow \text{Output}([\mathbf{h}_i; \mathbf{x}_{i:}])$
    end for
    for each region i do                    ▷ Location-specific regression
        $\hat{y}_i \leftarrow$ Linear Regression $(\mathbf{x}_{i:};\ \hat{s}_i)$
    end for
    $\Delta\mathcal{L}(\Theta) \leftarrow \text{BackProp}(\mathcal{L}(\Theta), \mathbf{y}, \hat{\mathbf{y}}, \Theta)$
    $\Theta \leftarrow \Theta - \eta\Delta\mathcal{L}(\Theta)$
end for
```

- **England** This dataset contains daily COVID-19 confirmed cases from 48 regions in England, ranging from March 13, 2020 to May 12, 2020. We consider this dataset as COVID-19 forecasting at very early stage. Locations are represented as the NUTS3 regions. The spatial connectedness matrix is calculated based on the mobility between regions, which is collected from the movement data of meta Data For Good disease prevention maps [37].
- **USA** This dataset contains daily COVID-19 confirmed cases from 60 counties in the US, ranging from September 1, 2020 to December 31, 2021. This dataset contains three most populated counties in the US (i.e. Kings county in New York, Cook county in Chicago and Los Angeles county in Los Angeles) and their nearby counties, ranging from September 2020 to December 2021. We consider this dataset for COVID-19 forecasting in a longer period. Locations are represented as the GADM level 2 regions. The spatial connected matrix is calculated from the social connectedness dataset of meta data for good project [38].

We collected county-level search data from Google COVID-19 search trends symptoms dataset [8]. Specifically, according to the existing publications, we consider several symptoms, i.e. 'fever', 'cough', 'hay fever', 'fatigue', 'diarrhea', 'rhinitis' and 'shortness of breath'. For England, we track five symptoms including 'fever', 'cough', 'hay fever', 'rhinitis' and 'shortness of breath'. For the US, we also track 'fever', 'cough', 'hay fever', 'fatigue' and 'diarrhea' because 'rhinitis' and 'shortness of breath' volumes are missing for US counties. Search volumes of each symptom is normalized to 0-100 as the normalized popularity of a symptom.

B. Experimental Setup

We train the models using the data from day 1 to day T to predict disease incidence at day $T + k$. [39] reports that certain search symptoms (e.g. 'fever') could reach the highest prediction performance when k is equal or larger than 5. Therefore, we set k from 1 to 7 days in this study. For England, we increase T one day at a time with T initially set to 30 days and a validation set of last 10 days. For the US, we increase T one month at a time with T initially set to 2 months and validation set of last one month. With this experiment setup, we can predict disease incidence as early as possible.

We evaluate the performance of the models using the mean absolute error (MAE) since absolute changes in the disease cases are most widely-used metrics in pandemic forecasting task:

$$MAE = \frac{1}{n} \sum_{u \in V} \left| \hat{y}_u^{(t)} - y_u^{(t)} \right| \tag{6}$$

Note that all reported cases are normalized with the population of that region throughout the experiments (i.e. cases per million people).

C. Baselines

We compare our model with several state-of-the-art methods as listed below:

- **Autoregressive Moving Average (ARMA)** [40] represents the linear autoregressive model. ARMA contains the autoregressive terms and moving-average terms together. The order of the moving average is set to 2 in implementation.
- **Random Forest (RF)** [41] is a non-linear regression model, which is a meta estimator that fits a number of regression decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

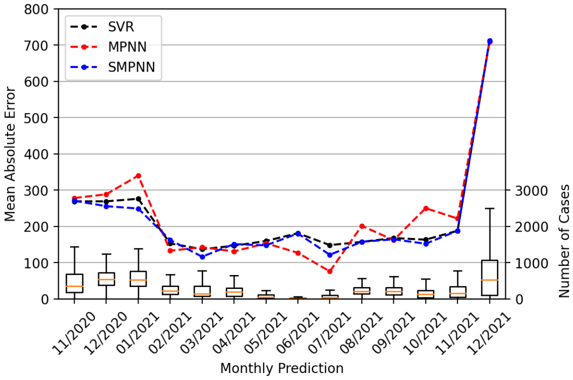


Fig. 4. Monthly predictions for the US.

TABLE IV
THE SETTING OF HYPERPARAMETERS

Hyperparameter	Value
Maximum number of epochs	100
Initial learning rate	$1e-3$
Batch size	32
Dropout rate	0.5
Feature window/days	2
Graph window/days	7
Early stop epochs	50
Validation days for UK	10
Validation days for USA	30

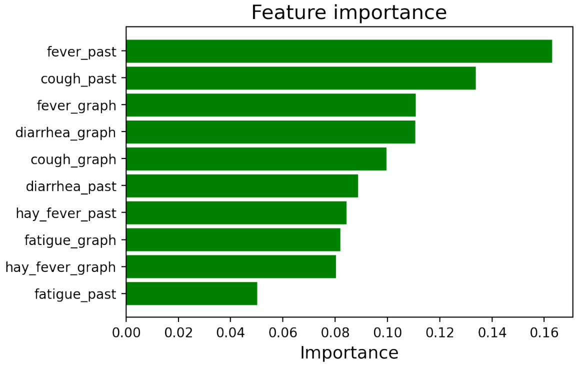


Fig. 5. Feature importance for SMPNN (k=7).

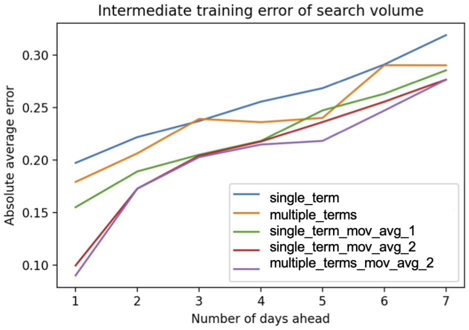


Fig. 6. Intermediate training errors when training SMPNN model.

- **Support Vector Regression (SVR)** [42] is a non-linear regression model, which is a nonparametric technique which relies on kernel functions to make predictions.
- **MPNN** [16] by design, could serve as an end-to-end model to predict the disease incidence from the search activity graph. Comparing to the location-specific regression models, we follow a similar design as described in section III-B1 while replacing the loss function as below:

$$\mathcal{L} = \frac{1}{n} \sum_{u \in V} \left(y_u^{(t+k)} - \hat{y}_u^{(t+k)} \right)^2 \tag{7}$$

where y_u^{t+k} denotes the reported number of cases for region u at day $t+k$ and $\hat{y}_u^{(t+k)}$ denotes the predicted number of cases.

Hyper-parameter Setting For all the models, we use the same validation set to select the best model as decribed in section IV-B. Specifically, for RF model, we explore the tree depth from 3 to 9 to control model complexity. For SVR model, we use polynomial kernel and explore the regularization term C from 0.1 to 2. For all the neural network models, as shown in Table IV, we use two neighborhood aggregation layers with the number of hidden units equals to 64 and store the model that achieves highest validation accuracy. To control model complexity, we apply batch normalization and dropout with ratio set to 0.5 to every neighborhood aggregation layer.

D. *Prediction Performance*

Table III summarizes the comparison between SMPNN and baselines for the pandemic forecasting tasks on England and US datasets. We investigate the different settings of predicting disease incidence one to seven days ahead ($k = 1, 2, \ldots, 7$). For the regression tasks, we report the mean absolute error (MAE) of disease cases for two countries. For England, the models are trained and predict daily in a two-month window. SMPNN outperforms all the baseline methods in 6/7 tasks, with MAE reduction up to 5.0% when k equals to 6. For the US, the models are trained the predict daily in a sixteen-month window. SMPNN outperforms all the baseline methods in 6/7 tasks, with MAE reduction up to 6.9% when k equals to 5. For England and the US, the lowest MAE is achieved when k equals 6 and 7 respectively, which is consistent with previous studies [5], [39].

E. *Contribution of Models and Features*

The baseline ARMA, RF and SVR models rely on location-specific dynamics for training, while end-to-end MPNN replies on cross-location dynamics with graph as the input. By design, SMPNN learns from both location-specific and cross-location dynamics, thus achieving lowest prediction errors as shown in in Table III. We further investigate how different models perform at different stages after disease outbreaks. As shown in Fig. 4, the box plots show the distribution of monthly new COVID-19 cases and the line plots represent the mean

TABLE V
PEARSON CORRELATION OF TOP TEN SEARCH TERMS FOR UK AND USA ACROSS ALL REGIONS.

Search Term in UK	**Correlation**	**Search Term in USA**	**Correlation**
Rhinitis	0.446^{**}	Ageusia	0.640^{***}
Hay fever	0.442^{**}	Anosmia	0.604^{***}
Hair loss	0.390^{*}	Low grade fever	0.548^{***}
Allergy	0.366^{*}	Fever	0.527^{***}
Abdominal obesity	0.362^{*}	Pneumonia	0.501^{***}
Dermatitis	0.359^{*}	Hypoxemia	0.468^{***}
Itch	0.358^{*}	Chills	0.466^{***}
Sleep disorder	0.347^{*}	Common cold	0.459^{***}
Rosacea	0.305^{*}	Shivering	0.416^{**}
Insomnia	0.297	Dysgeusia	0.409^{**}

Notes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

TABLE VI
PEARSON CORRELATION OF BOTTOM TEN SEARCH TERMS FOR UK AND USA ACROSS ALL REGIONS.

Search Term in UK	**Correlation**	**Search Term in USA**	**Correlation**
Pericarditis	0.003	Myalgia	0.172
Tumor	0.003	Xerostomia	0.166
Rheum	0.002	Infection	0.164
Bunion	0.002	Erectile dysfunction	0.151
Ataxia	0.002	Hypochondriasis	0.151
Anemia	0.002	Grandiosity	0.137
Petechia	0.001	Bradycardia	0.136
Blushing	0.0003	Periorbital puffiness	0.136
Varicose veins	0.0001	Burning chest pain	0.130
Hypoglycemia	0.00006	Palpitations	0.127

Notes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

absolute error for SVR, MPNN and SMPNN models. At the early stage of prediction (i.e. the earliest predictions in 11/2020, 12/2020 and 01/2021), SMPNN model outperforms all other models. SMPNN and SVR achieve the lowest MAE in 03/2021 while MPNN achieves the lowest MAE four months later in 07/2021. We also investigate how the location-specific features and the graph-generated features contribute to SMPNN model by calculating their average weights. As shown in Fig. 5, 'fever_past' and 'cough_past' (location-specific features) contribute most to SMPNN model, which is consistent with previous studies where search relevant to "fever" and "cough" contributes most to COVID-19 prediction [6], [39]. 'fever_graph', 'diarrhea_graph' and 'cough_graph' (graph-generated features) are ranked third to fifth out of ten features, which shows cross-location dynamics is also important for COVID-19 forecasting.

F. Search Terms Analysis

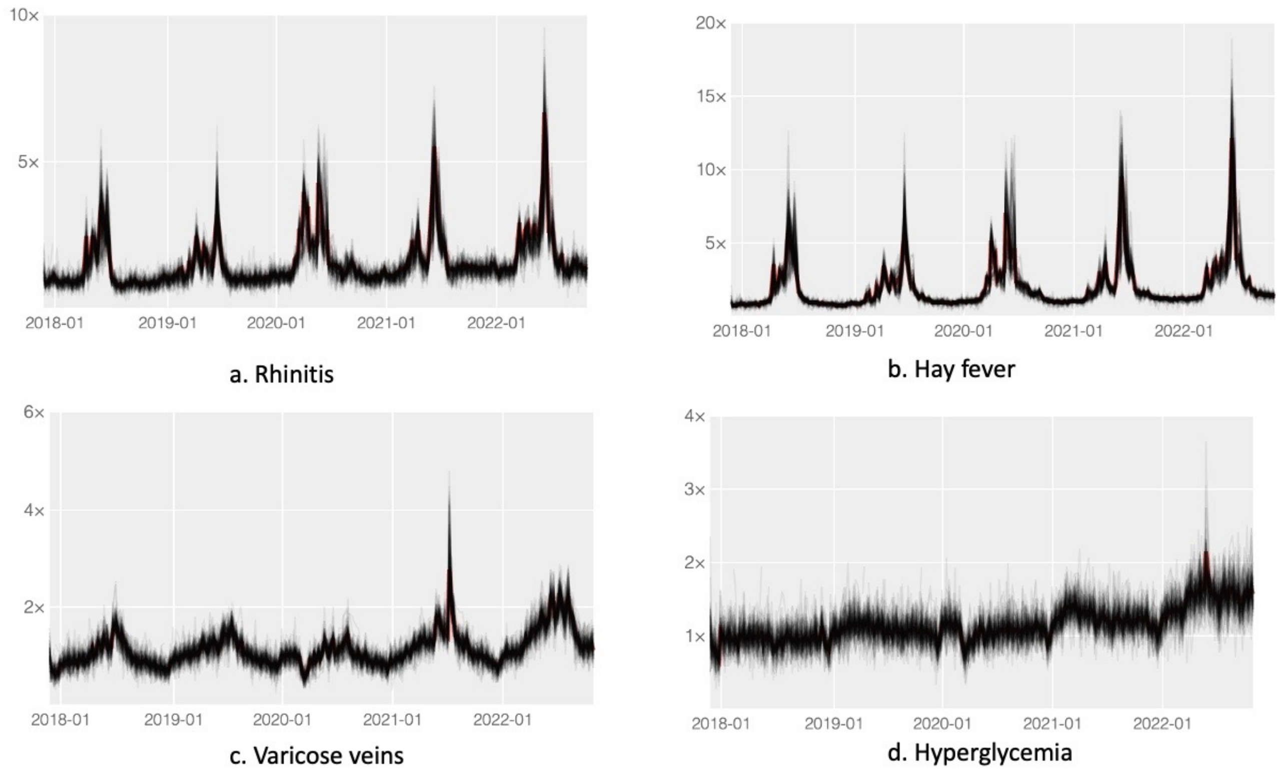
Aside from the predictive capability of the model, we also explored the impact of the search terms on the model's performance during training stage. Two perspectives have been taken into consideration when evaluating these terms. As a first step, we proposed several term combinations which would be fed into the model and training errors would then be measured. The dataset from the UK is used to measure training errors in our analysis. For single term, we used the 'rhinitis', and 'fever', 'cough', 'hay fever', 'rhinitis' and 'shortness of breath' five terms as multiple terms. Furthermore, we also compared these different data representations (i.e., utilizing the moving average in our case) at the same time. As shown in Fig. 6, it's clear that using only one of the terms (i.e. 'rhinitis') could introduce the most training error during the training phase. The training error decreases as the number of terms increases, which could be interpreted as evidence that COVID-19 involves a wide variety of symptoms. A moving average is more likely to provide a lower training error when compared to a time series representation.

In addition, we compared the Pearson correlation between the top ten and lowest ten search terms in the UK and the USA. From Table V and VI, it can be seen that search terms vary between countries, which highlights the importance of location-specific regression. It has been found that search terms with a high Pearson correlation are relevant to the symptoms of COVID-19 within a country. In the future, we will also explore the potential of these terms and leverage them as part of our work.

Furthermore, to validate our model, we also visualize the search term trends within UK and USA in Fig. 7 and Fig. 8. This COVID-19 Search Trends Symptoms dataset [43] provides aggregated, anonymous trends in the Google searches for over 400 health symptoms, signs, and conditions, such as cough, fever, difficulty breathing, and other health conditions that are commonly searched for online. For each region, the dataset gives a time series of the number of searches that have been conducted for each of the symptoms over time. These charts about symptom searches in the United Kingdom could display various types of data related to the frequency and distribution of online searches for COVID-19 symptoms across different regions in the country. We can observe that the conditions 'Rhinitis', 'Hay fever' for UK and 'Ageusia', 'Anosmia' for USA contribute a higher frequency with time, which aligns with our model's use. We also observe 'Varicose veins', 'Hyperglycemia' for UK and 'Burning chest pains', 'Palpitations' for USA contribute lower frequency with time, which aligns with our observations with their low Pearson correlation with COVID-19 cases. Fig. 7 and Fig. 8 show the search terms trend along with time. On these charts, the peak indicates that there have been more searches related to the search term. According to the search trends across all sub-regions, we observe that they share a similar trend during the progress of COVID-19, which validates our design to include geographical proximity information in our model design.

V. CONCLUSION

In this paper, a novel approach to pandemic forecasting is introduced, combining web search activity data and location relationships in a graph. The proposed framework, SMPNN, merges the best of existing message passing networks and location-based regression models. The method was validated using two real-world COVID-19 datasets and was shown to outperform prior state-of-the-art models, particularly in the early stages of outbreaks, by incorporating spatial graph features. This work makes significant advancements in the field of disease surveillance and forecasting, offering a new approach, methodology, datasets, and insights that integrate web search data and spatial information.



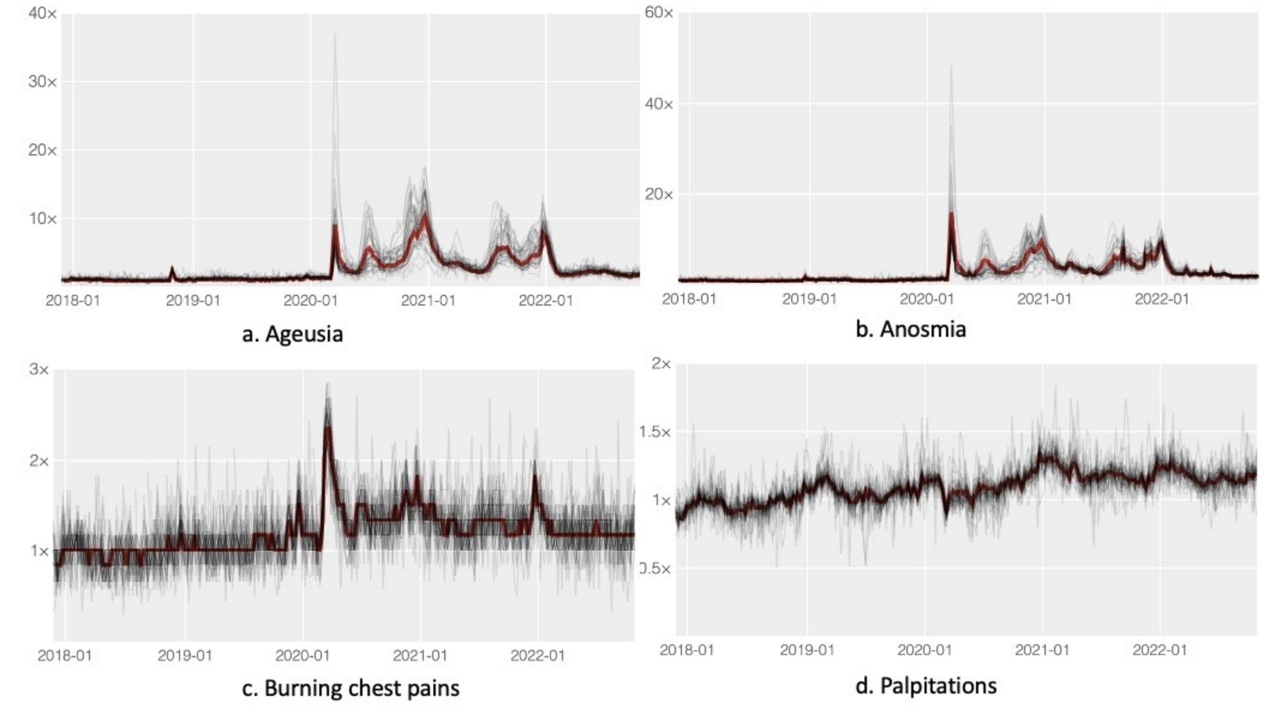
a. Rhinitis

b. Hay fever

c. Varicose veins

d. Hyperglycemia

Fig. 7. Normalized search volume ratios for search terms across all sub-regions in the UK (COVID-19 Search Trends symptom dataset [43]).



a. Ageusia

b. Anosmia

c. Burning chest pains

d. Palpitations

Fig. 8. Normalized search volume ratios for search terms across all sub-regions in the USA (COVID-19 Search Trends symptom dataset [43]).

Acknowledgment

This work was partially supported by grants from the US National Institutes of Health (NIH) National Library of Medicine (R21LM013014). The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

REFERENCES

[1] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[2] J. Zhou, "Improving interactive search with user feedback," Ph.D. dissertation, Emory University, 2022.

[3] C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *PLoS ONE*, vol. 5, no. 11, p. e14118, 2010.

[4] A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, G. Karypis, T. Q. Tran, and J. Sakuma, "Seasonal-adjustment Based Feature Selection Method for Predicting Epidemic with Large-scale Search Engine Logs," *arXiv*, pp. 2857–2866, 2019.

[5] V. Lampos, M. S. Majumder, E. Yom-Tov, M. Edelstein, S. Moura, Y. Hamada, M. X. Rangaka, R. A. McKendry, and I. J. Cox, "Tracking COVID-19 using online search," *npj Digital Medicine*, vol. 4, no. 1, p. 17, 2021.

[6] B. A. Panuganti, A. Jafari, B. MacDonald, and A. S. DeConde, "Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter," *Otolaryngology–Head and Neck Surgery*, vol. 163, no. 3, pp. 491–497, 2020.

[7] E. Yom-Tov, V. Lampos, T. Inns, I. J. Cox, and M. Edelstein, "Providing early indication of regional anomalies in COVID-19 case counts in England using search engine queries," *Scientific Reports*, vol. 12, no. 1, p. 2373, 2022.

[8] G. LLC. (2022) Google covid-19 search trends symptoms dataset. [Online]. Available: http://goo.gle/covid19symptomdataset

[9] P. Copeland, R. Romano, T. Zhang, G. Hecht, D. Zigmond, and C. Stefansen, "Google disease trends: An update," in *International Society of Neglected Tropical Diseases 2013*, 2013, p. 3.

[10] B. Zou, V. Lampos, and I. Cox, "Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data," 2019, pp. 2505—2516.

[11] J. Zhou, E. Agichtein, and S. Kallumadi, "Diversifying multi-aspect search results using simpson's diversity index," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2345–2348.

[12] J. Zhou and E. Agichtein, "Rlirank: Learning to rank with reinforcement learning for dynamic search," in *Proceedings of The Web Conference 2020*, 2020, pp. 2842–2848.

[13] J. Zhou, S. M. Zahiri, S. Hughes, K. Al Jadda, S. Kallumadi, and E. Agichtein, "De-biased modeling of search click behavior with reinforcement learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1637–1641.

[14] J. Zhou, S. Zahiri, S. Hughes, S. Kallumadi, K. Al Jadda, and E. Agichtein, "User click modelling in search queries," May 5 2022, uS Patent App. 17/514,522.

[15] S. Deng, S. Wang, H. Rangwala, L. Wang, and Y. Ning, "Graph Message Passing with Cross-location Attentions for Long-term ILI Prediction," *arXiv*, 2019.

[16] G. Panagopoulos, G. Nikolentzos, and M. Vazirgiannis, "Transfer Graph Neural Networks for Pandemic Forecasting," *arXiv*, 2020.

[17] J. Zhou, J. Ni, and Y. Rao, "Block-based convolutional neural network for image forgery detection," in *Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings 16*. Springer, 2017, pp. 65–76.

[18] C. Fritz, E. Dorigatti, and D. Rügamer, "Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly covid-19 cases in germany," *Scientific Reports*, vol. 12, no. 1, p. 3930, 2022.

[19] L. Wang, A. Adiga, J. Chen, A. Sadilek, S. Venkatramanan, and M. Marathe, "Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, 2022, pp. 12 191–12 199.

[20] H. A. Carneiro and E. Mylonakis, "Google trends: a web-based tool for real-time surveillance of disease outbreaks," *Clinical infectious diseases*, vol. 49, no. 10, pp. 1557–1564, 2009.

[21] Y. Teng, D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, and Y. Tong, "Dynamic forecasting of zika epidemics using google trends," *PloS one*, vol. 12, no. 1, p. e0165085, 2017.

[22] G. J. Milinovich, G. M. Williams, A. C. Clements, and W. Hu, "Internet-based surveillance systems for monitoring emerging infectious diseases," *The Lancet infectious diseases*, vol. 14, no. 2, pp. 160–168, 2014.

[23] C. Lin, S. Yousefi, E. Kahoro, P. Karisani, D. Liang, J. Sarnat, E. Agichtein *et al.*, "Detecting elevated air pollution levels by monitoring web search queries: Algorithm development and validation," *JMIR Formative Research*, vol. 6, no. 12, p. e23422, 2022.

[24] E. L. Aiken, A. T. Nguyen, and M. Santillana, "Towards the Use of Neural Networks for Influenza Prediction at Multiple Spatial Resolutions," *arXiv*, 2019.

[25] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of google flu: traps in big data analysis," *science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[26] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen, "Advances in nowcasting influenza-like illness rates using search query logs," *Scientific reports*, vol. 5, no. 1, pp. 1–10, 2015.

[27] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," *PloS one*, vol. 6, no. 5, p. e19467, 2011.

[28] L. Shi, G. Song, G. Cheng, and X. Liu, "A user-based aggregation topic model for understanding user's preference and intention in social network," *Neurocomputing*, vol. 413, pp. 1–13, 2020.

[29] Z. Wang, P. Chakraborty, S. R. Mekaru, J. S. Brownstein, J. Ye, and N. Ramakrishnan, "Dynamic poisson autoregression for influenza-like-illness case count prediction," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1285–1294.

[30] M. Won, M. Marques-Pita, C. Louro, and J. Gonçalves-Sá, "Early and real-time detection of seasonal influenza onset," *PLoS computational biology*, vol. 13, no. 2, p. e1005330, 2017.

[31] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks. 2017," *ArXiv abs/1609.02907*, 2017.

[32] F. Xie, Z. Zhang, L. Li, B. Zhou, and Y. Tan, "EpiGNN: Exploring Spatial Transmission with Graph Neural Network for Regional Epidemic Forecasting," *arXiv*, 2022.

[33] A. Tomy, M. Razzanelli, F. Di Lauro, D. Rus, and C. Della Santina, "Estimating the state of epidemics spreading with graph neural networks," *Nonlinear Dynamics*, vol. 109, no. 1, pp. 249–263, 2022.

[34] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," *arXiv preprint arXiv:1809.10341*, 2018.

[35] J. Zhu, M. Hong, R. Du, and H. Li, "Alleviating neighbor bias: augmenting graph self-supervise learning with structural equivalent positive samples," *arXiv preprint arXiv:2212.04365*, 2022.

[36] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.

[37] A. Herdağdelen, A. Dow, S. Bogdan, M. Payman, and A. Pompe, "Protecting privacy in facebook mobility data during the covid-19 response," *Facebook Research*, 2020.

[38] M. Bailey, R. Cao, T. Kuchler, J. Stroebel, and A. Wong, "Social Connectedness: Measurement, Determinants, and Effects," *Journal of Economic Perspectives*, vol. 32, no. 3, pp. 259–280, 2018.

[39] E. Yom-Tov, V. Lampos, T. Inns, I. J. Cox, and M. Edelstein, "Providing early indication of regional anomalies in COVID-19 case counts in England using search engine queries," *Scientific Reports*, vol. 12, no. 1, p. 2373, 2022.

[40] T. Kufel *et al.*, "Arima-based forecasting of the dynamics of confirmed covid-19 cases for selected european countries," *Equilibrium. Quarterly Journal of Economics and Economic Policy*, vol. 15, no. 2, pp. 181–204, 2020.

[41] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[42] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[43] G. LLC. (2022) Explore covid-19 symptoms search trends. [Online]. Available: https://pair-code.github.io/covid19_symptom_dataset