IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDP7020

Publicized:2024/08/29

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Integrating Cyber-Physical Modeling for Pandemic Surveillance: A Graph-Based Approach for Disease Hotspot Prediction and Public Awareness

Waqas NAWAZ[†a)], **Muhammad UZAIR**[††b)], **Kifayat ULLAH KHAN**[†††c)], ***and*** **Iram FATIMA**[††††d)], ***Nonmembers***

SUMMARY The study of the spread of pandemics, including COVID-19, is an emerging concern to promote self-care management through social distancing, using state-of-the-art tools and technologies. Existing technologies provide many opportunities to acquire and process large volumes of data to monitor user activities from various perspectives. However, determining disease hotspots remains an open challenge considering user activities and interactions; providing related recommendations to susceptible individuals requires attention. In this article, we propose an approach to determine disease hotspots by modeling users' activities from both cyber- and real-world spaces. Our approach uniquely connects cyber- and physical-world activities to predict hazardous regions. The availability of such an exciting data set is a non-trivial task; therefore, we produce the data set with much hard work and release it to the broader research community to facilitate further research findings. Once the data set is generated, we model it as a directed multi-attributed and weighted graph to apply classical machine learning and graph neural networks for prediction purposes. Our contribution includes mapping user events from cyber- and physical-world aspects, knowledge extraction, dataset generation, and reasoning at various levels. Within our unique graph model, numerous elements of lifestyle parameters are measured and processed to gain deep insight into a person's status. As a result, the proposed solution enables the authorities of any pandemic, such as COVID-19, to monitor and take measurable actions to prevent the spread of such a disease and keep the public informed of the probability of catching it.

key words: *Pandemic, COVID-19, Graph Neural Networks, Disease Hotspots, Social Mobility*

1. Introduction

Our world has recently witnessed the COVID-19 pandemic, which has affected many aspects of life, including human health, trade, mobility, and the economy [1]. Many countries around the world were affected, where almost 95 million people were reported to be COVID-19 positive; among them, 2.4 million are dead, and approximately 68 million recovered from this disease without any authentic vaccine [2]. Common symptoms among affected people include fever, coughing, shortness of breath, and loss of smell and taste, which can cause failure of different body organs such as the lungs and kidneys. It is most active and contagious in the first three days when a person shows symptoms [3]. Symptoms usually appear in the first five days, ranging from two to fourteen. The virus typically spreads among people through physical contact, coughing, breathing, and sneezing [3]. However, understanding the spread of such a disease is not trivial due to limited information about people's mobility and interaction patterns.

Experts and scientists are fighting to minimize the devastating impact of COVID-19 on humans and economy. Artificial intelligence (AI) experts are working on several fronts such as the processing of X-rays of patients with COVID-19 to determine differences from already existing diseases [4][5], monitoring and controlling its spread in various localities [3] [6] [7] [8] [9] [10], effects on mental health [11] [12] [13], and prediction and identification of infected areas [1][2][14][15][16][17][18] to name a few. Research is expanding due to data availability, and AI algorithms are data-hungry for accurate predictions and trend analysis. Information about COVID-19 in the cyber world from social networks is increasing rapidly while being unstructured and in various formats [6]. Data dynamics vary rapidly, and extracting the desired information is a challenging task [3]. The real-world facts associated with the disease, such as location, temperature, population, age distribution, number of infected and recovered cases, death rate, and similar [1], are easily available on various websites and newspapers in almost every region, that is, at the country level. However, extracting such data at the city or subregion level is difficult [7].

Despite the scarcity of fine-granular data, many researchers proposed techniques for the prediction and identification of hotspots using cyberworld information [3] [2] or completely relying on real-world facts [15] [10] [17]. For example, using NLP techniques, the authors of [3] used tweets to collect information on the location of people infected with COVID-19. Another study [2] uses mobile phone data to track people's mobility to predict mobility hotspots. In this way, they used real-world facts as features (such as geography, weather, and quantitative trends of COVID-19) for hotspot prediction using the K-nearest neighbor (KNN) ap-

[†]Department of Information Systems, College of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia
[††]Department of Computer Science, National University of Computer and Emerging Sciences (FAST), Islamabad, Pakistan
[†††]College of Accounting, Finance, and Economics, Birmingham City Business School, Birmingham City University, Birmingham, United Kingdom
[††††]Department of Computer Science, King Faisal University, Al-Hofuf, Saudi Arabia
a) E-mail: wnawaz@iu.edu.sa
b) E-mail: emyou42@gmail.com
c) E-mail: kifayat.khan@bcu.ac.uk
d) E-mail: iram.fa@gmail.com

proach. However, the authors in [15] and [10] statistically evaluated the geospatial characteristics extracted from real-world facts for disease hotspot analysis. The trends in distribution and the number of COVID-19 hotspot counties in the United States were identified in [17] by the Centers for Disease Control and Prevention (CDC) using factual data. In a relevant study [7], the authors identified dangerous spots using a time series forecasting model based on data from daily reported cases from government authorities. Most of these studies are based on facts from the real world, and these works lack identification of the origin of the spread [14], where knowing the location of infected people is a bottleneck. Therefore, accurately predicting COVID-19 hotspots is challenging when considering patient mobility and locality information where the number of cases is increasing.

To overcome the aforementioned issues, we present a novel approach to predict disease hotspots using data from the cyber and physical worlds with the help of graph convolutional networks. We understand that data about the disease from both worlds, i.e., cyber-world information and real-world facts, are complementary to present an enhanced and holistic view. We perform rigorous data preprocessing to aggregate and establish a relationship between them. The intention behind aggregating the information of both types is to get the best of both worlds, since some of the features available in the cyber world do not exist (or are hard to obtain) in the data of the physical world and vice versa. We model data in the form of a directed multi-attributed weighted graph. In our modeling, a node represents a location, whereas an edge and its direction represent the mobility of people between two regions. The additional information of a particular locality (e.g., reported cases, temperature, population density, etc.) is associated with the corresponding node in our graph. The resultant snapshot of the graph database serves as an input to Graph Convolutional Networks (GCN) for the prediction of hotspots. The significant contributions of this paper are as follows.

- We acquire and pre-process the raw data from the cyber and physical worlds to not only analyze the mobility patterns of people in a particular region but also understand the spread of disease through COVID-19 reported cases in that region.
- We release our data set to serve the wider research community, enabling more versatile and extensive research.
- We present a unique data modeling strategy to aggregate and integrate logically related COVID-19 data to predict hotspots. Our approach initially aggregates data for cyber-world information and real-world facts at a particular granularity level. Then, it integrates it as a directed multi-attributed weighted graph, where locality information plays a pivotal role in this process. The motivation is to collect detailed information about our subject of interest since data from a single source do not contain all the required information.
- We predict the hotspots in a particular region at fine granularity (e.g., subregion or subdistrict) using classical machine learning and graph-convolutional networks (GCN). In contrast to existing approaches, our trained models identify possible infected subregions even in the absence of previously reported cases. Based on mobility patterns, GCN helps propagate the effect of spreading disease to neighboring regions.

2. Related Works

Identifying the infected region is essential in controlling the spread of diseases such as Coronavirus (COVID-19). The information required for the above-mentioned purpose is crucial for an effective strategy. Therefore, in this section, we highlight existing studies in the literature related to acquiring and modeling helpful information and different vital strategies to use this information in the detection and mobility recommendations of hazards for susceptible individuals.

2.1 COVID-19 Data Acquisition and Modeling

Information related to COVID-19 is generally complex and difficult to obtain due to limited availability and privacy concerns. However, researchers have used public platforms such as the Web and social networks such as Twitter (now known as X) and Facebook to obtain related information, which is not trivial and requires an effective strategy. It is expected to obtain the stream of filtered tweets and preprocess those tweets into a consolidated valuable information repository of COVID-19 [19] [20] [21]. The authors in [19] proposed 900 different keywords related to COVID-19 to fetch tweets from the Twitter repository using different APIs such as Tweepy and various search APIs. The data fetched had 524 million records of multilingual tweets spanning 90 days, most of which were in English. Each tweet contains the tweeted text, geo-coordinates, and user account information, whereas most tweets lack location information. The authors used the gazetteer-based approach [19] in the name entity recognition system to determine the location information from the tweet's text. The accuracy of the tweet's location is still low because the tweet content mostly talks about other locations while accumulating noisy data based on 900 keywords.

Similarly, the authors in [20][21] prepared a COVID-19 data set based on Twitter tweets for the research community. In [20], they used the tweepy API, the streaming API, and the search API with different keywords to collect past multilingual tweets related to COVID-19. They fetched approximately 123 million tweets, of which almost 60% were in English. The authors aimed to extract vital information from tweets, such as finding areas where COVID-19 patients are increasing in number and their spread according to available location information. The approach used resulted in gathering tweets with rumors, talks, or assumptions.

On a similar note, the authors in [22][23] collect, analyze and track COVID-19 trends in different regions of the world. They developed an Amazon Web service-based

tracker[22], which collects data into data lake (that can hold structured, semi-structured, unstructured, and binary data) and displays them through dashboards. The data is automatically updated every 15 minutes using a Python script. They have used the Susceptible Exposed Infectious Recovered (SEIR) predictive model to show the trend of COVID-19 within and outside China. Visual representation of the data collected from authentic repositories dramatically helps to understand the trends of COVID-19. Identifying hazard areas of such a visual analysis at the acceptable level is challenging due to inaccessibility or limited knowledge of the location information. Data aggregation to preserve user privacy is one of the root causes of this restricted analysis. For this purpose, the authors in [23] use aggregated data from the Facebook social network to understand the spread of COVID-19 based on social interactions. The data suggest the regions with more social ties are prone to be hotspots since the social network has these relationships based on the aggregated movement of people from one region to another, the population density, and estimated income, hence useful to predict the spread of COVID-19.

Information about the reported cases of COVID-19 patients [24], also known as real-world facts, plays a crucial role in understanding and controlling the spread of this disease by authorities around the world. Therefore, many local (e.g., KCDC [25], NIH [26]) and international health organizations (e.g., WHO [27]) have maintained aggregated information on infection-reported cases for various purposes, which is publicly accessible on different platforms such as daily infection reports [28] [29][30][31]. The research community has also used additional information, such as population density[32], mobility [33][34] and weather forecast [35] to improve the precision of their proposed systems to predict possible infectious areas [1].

2.2 Predictions and Trend Analysis of COVID-19 using Machine Learning

Recommending places of interest to travelers based on friendships on location-based social networks is a useful approach [36][37]. However, determining whether a place to visit is safe is challenging, especially during any pandemic. In this regard, [1] developed a warning system to predict the hazardous areas where there is a chance of an outbreak of viruses. The authors scrapped data from various authentic sources and used KNN's classification to identify hazardous areas. They extracted features from data based on geographical, demographic, and temporal features. The geographical area is divided from 1 to 5, where 1 to 3 have a lower chance of an outbreak, while the areas with scales 4 and 5 are the main hotspots. The formulation task involves labeling the specific location l with associated characteristics x as "in hazard" (y=1) or "not in hazard" (y=0) as a simple binary classification. According to this study, demographic and geographical characteristics have stronger contributions because they are closely related to disease transmission.

The authors of [15] statistically evaluated geospatial characteristics to understand the spread and dynamics of the disease for hotspot analysis. The authors modeled the vulnerability zoning of COVID-19 using the weighted sum method based on the analytical hierarchy process (AHP-WSM) approach. Spatial-temporal hotspots were analyzed based on reported cases. The zone of vulnerability to death was determined based on several factors. The authors of another similar study [10] identified areas with high intensity of virus spread using heatmaps. They used one of the well-known statistical measures, the Moran index, to estimate spatial association or spatial autocorrelation among regions and cluster them to identify possible hotspots. The logistic growth model was also studied with the Moran index to understand the spread better. Data utilized for this purpose include reported cases captured at the national and provincial level by a research group at the University of Pretoria. This study also used demographic data for the population at the provincial level, such as the number of recoveries and deaths provided by the National Institute of Communicable Diseases (NICD). The CDC report [17] also provides trends in the distribution and number of COVID-19 hotspot counties in the US. Spatial statistical analysis was performed for similar diseases such as Malaria [16] and Influenza [18] to understand the distribution patterns of the disease and identify hotspots based on the collected data.

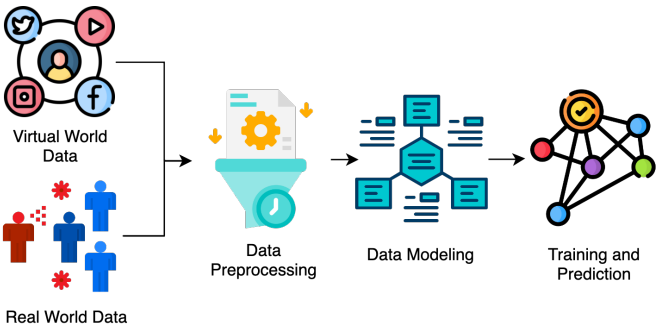
Predicting the spread of COVID-19 without systematic analysis of temporal data is also very interesting but non-trivial too. Therefore, the authors in [3] predicted the spread of COVID-19 by designing and analyzing different deep learning models using time series deep neural networks and the stacking ensemble technique. For better accuracy, they collected data from various sources, including mobility patterns, and used cluster-based training to overcome the sparsity problem. Their approach can only forecast every week at the country level due to a lack of information of finer granularity. In another study [2], the authors presented a graph neural network (GNN) based approach to predict the number of cases infected with COVID-19. In this approach, data at the country level is represented as large spatio-temporal graphs where nodes reflect the mobility of humans from one place to another, and the edges show inter-region connectivity. In [8], the authors designed a system to predict the situation of the next two days based on digital traces with mechanical models. They used data from different sources, including Chinese CDC reports, Internet searches, news and GLEAM, a daily coronavirus forecaster. The authors in [14] claim that people's mobility can alone dictate the risks involved in visiting a particular city; however, the availability of such data is scarce and makes it an difficult. [9] developed an AI-inspired technique to predict Coronavirus's size, length, and end time. The authors developed modified stacked autoencoders to be applied in confirmed cases in real time, where data is collected in two months from the World Health Organization (WHO) [27]. However, it predicted the epidemic's end in April 2020, which was inaccurate due to limited training data.

Above are some of the most relevant studies to our work,

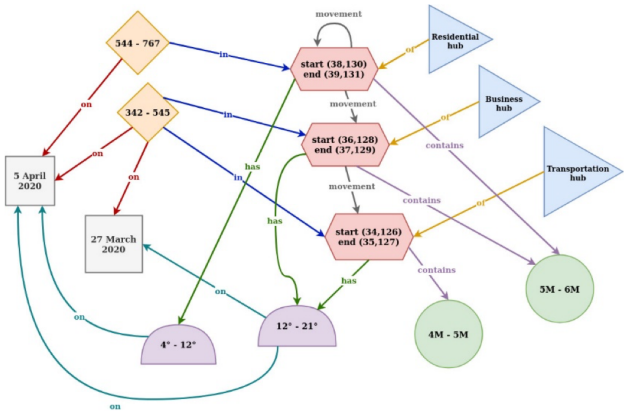
and we have provided a brief overview of them to present a broader picture. Below, we further present the summary of selected related works concerning our proposed approach in Table 1.

Table 1: Summary of the Related Works

Article	Virtual data	Real world Facts	Modeling	Granularity
[1]	No	Yes	No	City
[3]	No	Yes	No	Country
[2]	No	Yes	Yes	Country
[14]	Yes	No	No	City
[8] [9]	No	Yes	No	Province
[15]	No	Yes	No	District
[10]	No	Yes	No	Province
[16] [18]	No	Yes	No	City
[17]	No	Yes	No	County
Proposed	Yes	Yes	Yes	Subregions



(a) Keys stages to solve the problem



(b) Real world and cyber world data as a multi-attributed graph

Fig. 1: Overview of our proposed methodology.

3. Proposed Methodology

This section discusses our proposed approach for detecting disease hotspots. As the related work section shows, predicting hotspots related to a contagious disease such as COVID-19 in real life is not trivial due to the lack of relevant information on site. Therefore, in this study, we develop a strategy to establish an effective relationship between on-the-ground data (we call them real-world facts) and information from cyberspace social networks (i.e., Twitter, Facebook, etc. We use this relationship to predict disease-related hotspots as potential threat regions to help limit disease spread. In this section, we elaborate on data acquisition and preprocessing, data modeling as a disease-centric multi-attributed directed graph, and multi-attributed directed graph learning strategy using Graph Convolutional Network (GCN). Our proposed approach is illustrated in Figure 1, where data is acquired from cyberspace and physical world. The data is preprocessed and modeled in a multiattributed directed graph to establish a relationship between the virtual and physical worlds through locality information as shared attributes. Although location information is limited and insufficient in the real world, we uniquely leverage it with cyber-world data to overcome this issue. Moreover, we trained our graph data on GCN to extrapolate unknown areas in the real world toward hotspot detection with limited available information.

3.1 Data Acquisition and Pre-processing

The initial phase of our proposed strategy is to acquire and preprocess data from cyberspace and the physical world. Now, we elaborate on the data sources and our pre-processing strategies.

As the physical-world data, in our case, refers to COVID-19 reported cases as real-world facts obtained from the official Institute of Health in South Korea. We chose this region due to the availability of reported cases with sufficient granularity and cyberspace data. Their real-world facts data is publicly available on their official website. For cyberspace data, we consider Meta's initiative to collect vital information about communities, i.e., Data for Good [38]. We collect the aggregated data from Meta through the developer's account so South Korea can establish an effective relationship. The details of the datasets used in our approach are discussed in the Experimental Section.

In the pre-processing stage, we remove all unwanted information from raw data that has less potential to help identify hazardous areas. We then aggregate all the data into a single file using Panda's data frames. We integrate multidimensional data based on locality information, that is, latitude and longitude, and assign labels to 10% of the data based on assumptions. The rest of the data labels are determined through context-based filtering.

3.2 Proposed Graph Data Modeling Approach

To learn and consider mobility patterns between locations, we propose modeling data as a graph. We propose a unique strategy to model preprocessed data into a disease-centric multi-attributed graph. This data modeling approach enables us to extract valuable insights about people's mobility towards better hotspot identification. We present an example toy graph in Figure 2, where we represent each node with

geo coordinates and information, such as reported cases and population of that locality, as node attributes. Edges or links between nodes are created if people move between those coordinates.

To transform the pre-processed data into an aggregated graph structure, first, we use location and people count as mobility information from social networks to create the nodes. The data set contains only the end location, and the information from people moving to that location is an instance of the records. We manipulate that information to construct the source and destination locations and the number of people moving between those locations. This way, we create a link between those location nodes in the graph as a second step. The direction of the link specifies the direction of people's movement. Since data instances are separated daily, hence reported cases from COVID-19 are stored for each day. Therefore, we propose a scheme to aggregate daily information into a single graph instance.

3.3 Aggregation of multi-attributed directed weighted graphs

We aggregate all the scattered information obtained from each day's data into a single large graph, reflecting recent and previous insights about people's mobility and reported cases. In this way, a single large aggregated graph reflects recent and previous information. For this, we use weighted aggregation to generate a single graph from a set of graphs. Figure 3 demonstrates the said process on a toy graph.

As GCN takes only a single graph and trains a model on it, we ensure that all the information is provided in a single aggregated graph to identify the infected areas. Our proposed multi-attributed graph model contains nodes having attributes to incorporate the aforementioned information and the edges showing mobility information for a single day. This way, a date-wise timestamped graph is provided as input to the GCN for training and prediction purposes.

As recent information significantly impacts the prediction compared to older information. So, similar to the idea of Exponential Moving Average (EMA), a well-known statistical data analysis measure, we use the weighted aggregation in a way that if we have records of ten days, then the record of the first day has the least weight while the record of the tenth day carries the maximum weight [39]. In addition, we provide the following two different aggregation scenarios for the construction of our multi-attributed graph.

3.3.1 Scenario 1: Aggregating graphs having the same structure

In this scenario, we consider only the nodes along with their connectivity that exist from the first-day information of the collected dataset, i.e., using the same network structure for the entire period. Every node and edge stays in the resultant aggregated graph for the whole period. Weight information attached to each edge follows the same principle since the recent edges hold more impact than the older ones. In this way, we aggregate data of all the date ranges into a single aggregated graph, as demonstrated in Figure 3.

3.3.2 Scenario 2: Aggregated graphs with different structures

In contrast to the previous scenario, this type of aggregation is more realistic; as the network evolves over time, the addition and deletion of nodes and edges occur. Hence, we consider each graph's present shape to have as many new or existing nodes and edges as possible. In this way, each graph to be aggregated may possess a different network structure, but the resultant aggregated graph contains even a node or edge, which occurred only once during the entire lifetime of the data set collection period. Weight information is kept in the same manner as the previous scenario. We demonstrate the scenario using Figure 4.

3.4 Multi-Attributed Directed Graph Learning

After obtaining an aggregated multiattributed graph, reflecting the insights and impact of the past information using scenario 1 or scenario 2, we use it as input to GCN to predict the hotspots. Although the GCN operates similarly to any other classification algorithm, it is used to solve graph-specific prediction problems such as graph classification, node classification, link prediction, community detection, graph embedding, and graph generation, among others. In our case, we aim to identify whether a particular location is hazardous. In this way, it is a node classification problem for GCN [40] but for multi-attributed aggregated graph. We train the GCN using proposed graph-modeling approaches to predict infected areas as output.

Graph Convolutional Layers (GCLs) are the fundamental building blocks of Graph Convolutional Networks. They gather information from neighboring nodes and refine node features based on these collected data. Typically, a GCL performs a graph convolution operation, aggregating feature information from connected nodes and applying a neural network operation like a linear transformation and non-linear activation. In our model with two hidden layers, each layer acts as a GCL, processing and updating the input features by the underlying graph structure.

Fully Connected Layers (FCLs) refine the extracted features using standard neural network operations after the graph convolutional layers. Unlike localized connections in convolutional layers, FCLs establish connections between every neuron in one layer to every neuron in the next, enabling a comprehensive feature transformation. These layers prepare the learned features for the ultimate task, such as classification or regression.

Output Layer: The final layer of our GCN delivers predictions or classifications based on the features learned by the preceding layers. The GCN will identify and classify hazardous or nonhazardous areas in this application.

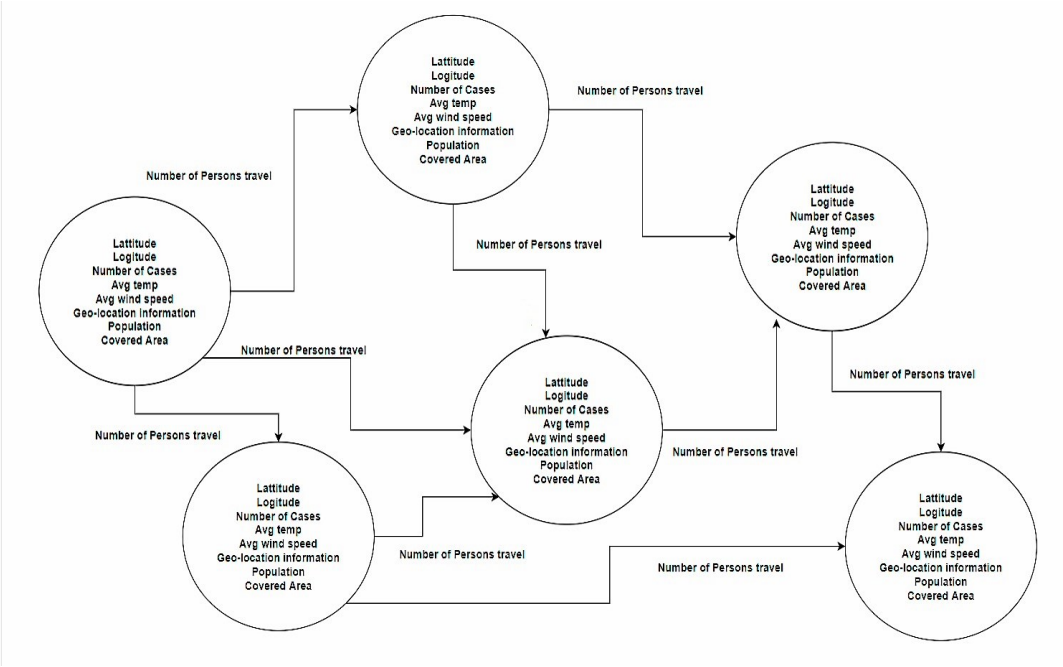


Fig. 2: Structure of a multi-attributed directed graph created using information from social networks and real-world facts as reported cases.

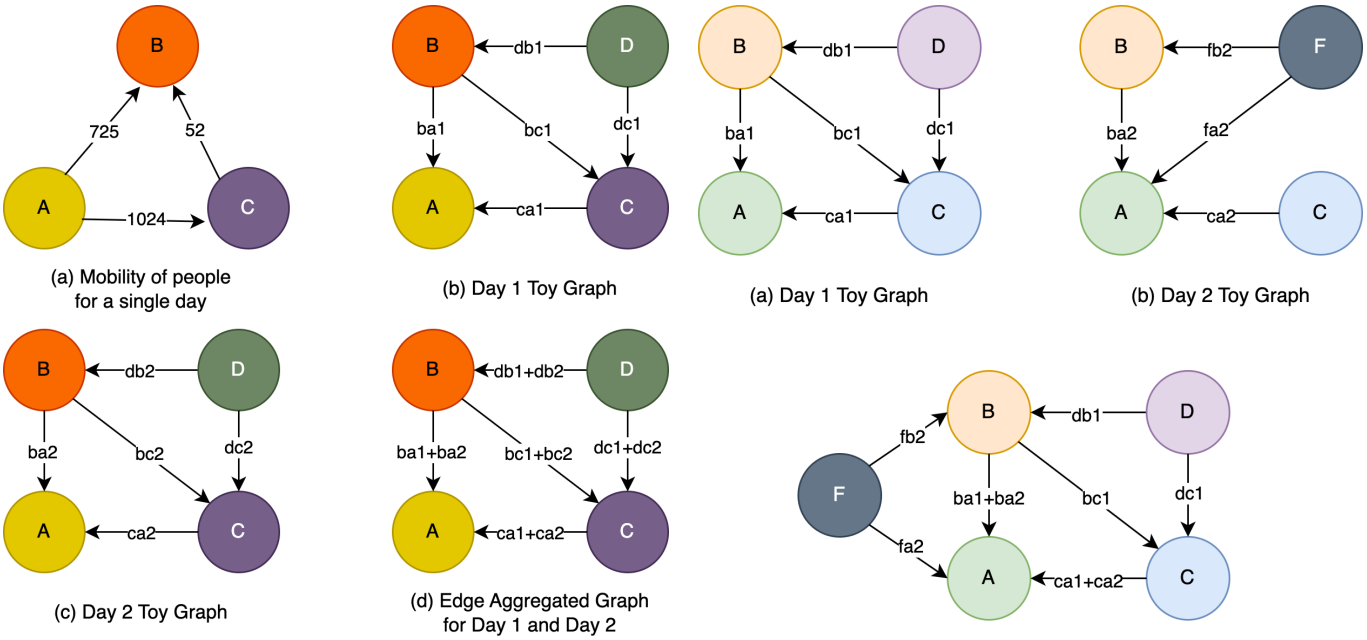


Fig. 3: Data aggregation scenario for the graphs having the same structures.

Fig. 4: Data aggregation scenario for graphs with different structures.

4. Experimental Evaluation

This section presents details of our experiments we performed on the generated data set. We evaluated the results using precision, F1 score, AUC score, and recall scores.

4.1 Details of Our Dataset

Our proposed solution effectively connects cyber and real-world facts to identify hazardous areas. The availability of such a dataset with cyber and real-world facts for each region

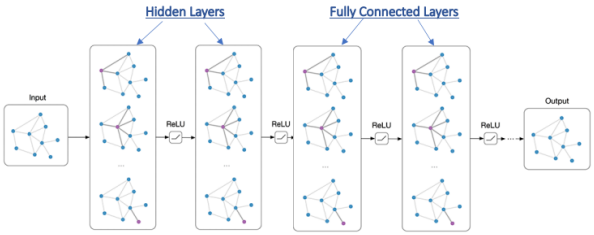


Fig. 5: Architecture of GCN

is challenging because users in cyberspace may not have their data in the physical world and vice versa. To create ease for further research in this domain and serve a more comprehensive research community, we released our data set of 180 days of South Korea with cyber-world and real-world facts. We understand that this aspect also serves as a useful contribution to our research. The following paragraphs explain our data collection procedure and its necessary details.

4.1.1 Twitter Dataset for Cyber World Facts

We selected tweets about COVID-19 having location information from various regions of South Korea. We used different hashtags to retrieve the required information and collected approximately 18000 tweets. Some of the useful attributes in these data are as follows:

- Created at: contains date and time, $MM : DD : YYYYHH : MM$, on which a tweet is created.
- Followers: This contains the number of followers of a user who tweets.
- Friends: This contains the number of friends of a user who tweets.
- Location: It is obtained using the geolocation option. This attribute is important to divide the tweets location-wise
- Text: Text or message from a tweet.
- Tweet ID: It is the ID of a tweet. It is useful to obtain further information from the raw tweets.

4.1.2 Facebook Dataset for Cyber World Facts

We collected Facebook data using a developer account. Facebook collects aggregated movement information for users from one place to another and presents the results to its audience using movement maps. Movement maps illustrate patterns of movement of groups of people between different neighborhoods or cities over several hours. These also tell the people's density of the regions via the active user's information. Some of the useful attributes include date & time (time represented by the current map layer), start location (a region where the movement of the group started), end location (a region where the movement of the group ended), length (distance traveled in kilometers), baseline movement of people (total number of people who moved from start to end location, on average, during the weeks before the disaster began), people movement during a crisis (total number of people who moved from start to end location during the specified time), the difference (difference between the number of people who moved from the starting location to the ending location during the disaster compared to before the disaster) and standard Z-score (number of standard deviations by which the count of people moving during the crisis differs from the number of people moving during the baseline. Any Z score greater than 4 or less than -4 is reduced to 4 or -4).

As this data contain information on the movement of people from one place to another during the COVID-19 pandemic, we show the trends derived by different linear regression models in Figure 6.

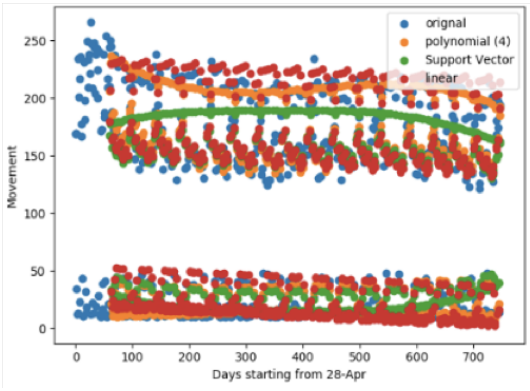
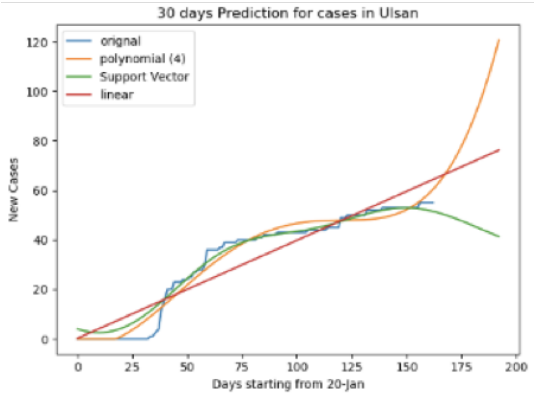


Fig. 6: Movement trends in different regions. Original refers to raw data on people's movement. We also fit various regression models to our data, including polynomial(4) of order 4, support vector, and linear regression.

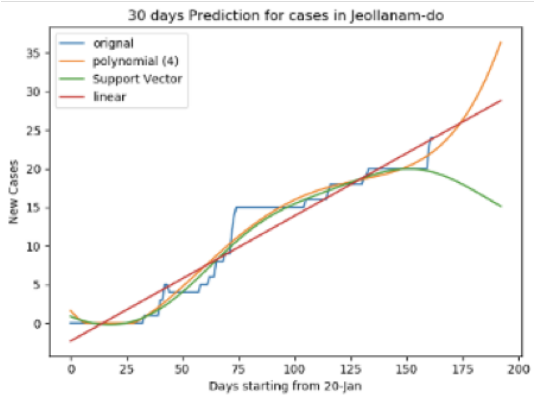
4.1.3 Real-World Facts

Our real-world facts come from the official website of KCDC (Korea Centers for Disease Control and Prevention) [†]. KCDC announces the information about COVID-19 quickly and transparently. The data set is divided into several files such as Cases (data for COVID-19 infection cases), epidemiological data of COVID-19 patients, time series data of COVID-19 status, location and statistical data of the regions, weather data of various regions, keyword search trends in NAVER (most popular search engine in South Korea), data of floating population in Seoul, South Korea (from SK Telecom Big Data Hub), and data of the government policy for COVID-19 in South Korea. After we get the real-world facts, we pre-process it to fetch only the required features suitable for our approach and apply a few linear regression models on it to understand the trends of the number of COVID-19 in different regions. We visualize the observed trends in Figure 7 that are helpful in further processing of the data to identify the hazard areas.

[†] https://www.kdca.go.kr/index.es?sid=a3



(a) Region: Ulsan



(b) Region: Jeollanam-do

Fig. 7: Trend of cases in different Regions.

4.1.4 Combining Cyber World and Physical World Facts

As soon as we finish preparing the data set for both domains, we establish an effective relationship between these two different independent domains for identifying hazardous areas. To achieve this, we join these two datasets based on Latitude and Longitude because these coordinates are the only common things in both datasets. This way, our dataset is created after several pre-processing steps.

4.2 Data Labeling for Ground Truth Establishment

Data labeling is one of the most important and important processes in any classification or regression problem such as ours to identify hazardous areas. In our data set, all attributes are independent, and their ranges are also different; that is, "Number of Confirmed cases" ranges from 0 to 1000 while the attribute "Temperature" ranges from -10 to 25. Therefore, we first normalize all attributes on a scale of 0 to 1. In this way, the lowest value of each attribute is 0, and the highest value is 1. We then assign weights to each attribute based on its importance in identifying hazardous areas. Attribute like "Confirmed Cases" is vital, so we assign the highest weight to it. Similarly, travel records of people from one place to another are also significant attributes because people traveling from hazardous areas to other places impact the spread of the disease. However, since COVID-19 is an airborne virus through tiny droplets and similar mediums [41], the wind speed has little impact on our target variable, so we assign the least weight to it.

Proposed-Phase 1	**Accuracy**	**F-Score**	**AUC Score**	**Precision**
KNN	94.59	0.90	0.57	0.91
SVM	79.15	0.59	0.48	0.50
GBDT	98.41	0.98	0.60	0.95
DT	75.97	0.39	0.54	0.33
NB	52.53	0.66	0.44	0.59
Proposed-Phase 2	**Accuracy**	**F-Score**	**AUC Score**	**Precision**
KNN	95.60	0.94	0.56	0.89
SVM	81.46	0.61	0.48	0.49
GBDT	98.20	0.98	0.57	0.96
DT	76.62	0.39	0.53	0.32
NB	52.16	0.66	0.44	0.60
Proposed-Phase 3	**Accuracy**	**F-Score**	**AUC Score**	**Precision**
KNN	94.95	0.91	0.56	0.87
SVM	79.51	0.58	0.47	0.50
GBDT	98.70	0.99	0.57	0.96
DT	76.26	0.39	0.53	0.33
NB	50.14	0.63	0.42	0.57
Previous Work[1]	**Accuracy**	**F-Score**	**AUC Score**	**Precision**
Phase1-GBDT	78	0.27	-	0.86
Phase2-GBDT	91.77	0.51	-	0.95
Phase3-GBDT	96.2	0.36	-	0.66

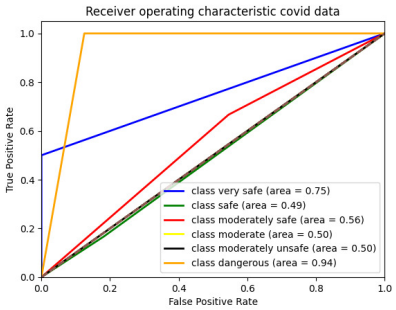
Table 2: Classification results on selected evaluation measures

	Precision	**Recall**	**F1-Score**	**Accuracy**	**Support**
Phase 1 GCN 50-20	0.51	0.50	0.46	0.59	17
Phase 2 GCN 100-50	0.51	0.47	0.42	0.65	8
Phase 3 GCN 150-150	0.86	0.79	0.83	0.8	17

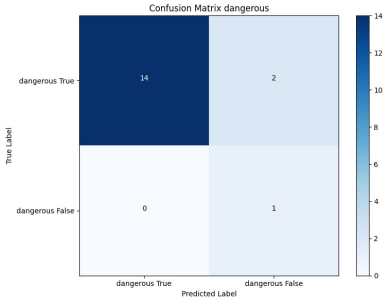
Table 3: GCN-based model prediction results on selected evaluation measures on different GCN hidden layer neurons settings.

Once we assign the weights to the attributes, we add all of them in the corresponding rows to get a single value, which serves as class labels for predicting the hazardous region. We then divide the data into different buckets to group the records and classify them from highly unsafe or hazardous to safe or non-hazardous. We also used context-based filtration to automatically label the data based on some sample data, that is, choose 10% of the data and label it based on its buckets. We use this algorithm to make the data labeling process efficient.

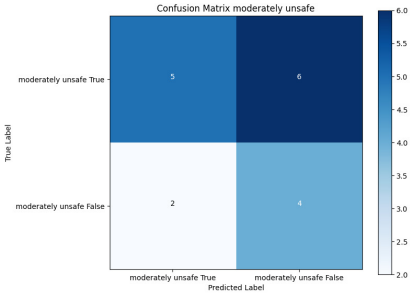
Existing approaches predominantly rely on reported cases representing physical world data. However, we propose that a more effective prediction model should incorporate additional factors such as mobility patterns and population density due to the disease's contagious nature. Integrating these relevant data points with the reported cases would provide a more comprehensive understanding and lead to more accurate predictions of hazardous areas.



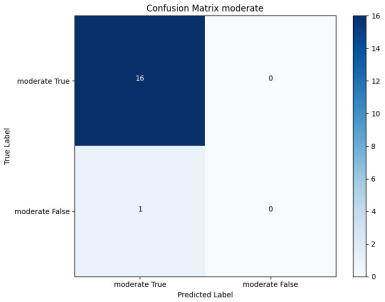
(a) ROC for All Classes



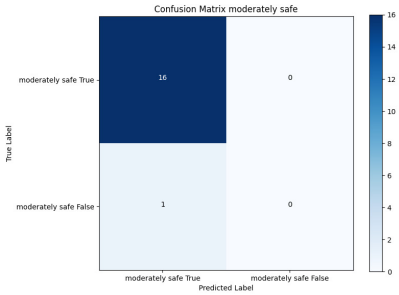
(b) Dangerous



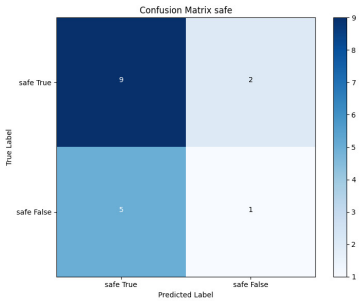
(c) Moderately Unsafe



(d) Moderate



(e) Moderately safe



(f) Safe

Fig. 8: ROC and Confusion Matrices.

To validate our hypothesis, we performed proof-of-concept experiments. In particular, we observed a significant increase in accuracy when considering both mobility and reported cases (64%) compared to using all available parameters (76%). Furthermore, we identified a specific scenario to demonstrate the effectiveness of integrated data. On 20th January 2020, four cases were reported in Incheon, South Korea. Within three days, six and two cases were reported in neighboring regions Gyeonggi-do and Gangwon-do respectively. Mobility data analysis revealed 448 movements from Incheon to Gyeonggi-do between January 20th and 24th, 2020. This finding illustrates the natural phenomenon of disease transmission through even limited movement between regions.

4.3 Classification Models Details

Once the data set is prepared, we apply classical machine learning models, i.e., K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradiant-Boosted Decision Trees (GBDT), Decision Tree, and Naive Bayes (NB). State-of-the-art research [1] in this domain also applies the same classification models, but only the data of real-world facts. Therefore, we understand that our data set is better than the one used in [1] because we include more versatile information to predict hot spots. We present the results obtained in our dataset in Table 2 and observe quite encouraging performance.

4.4 GCN Configuration and Model Predictions

In our initial experiments, we used a GCN model consisting of an input layer, two hidden layers, and an output layer. The input layer takes a multiattributed directed weighted graph as input. The hidden layers compute the required weights based on the structural patterns of the input graph. As we have to predict the hazardous areas, the prediction generated at the output layer includes one of the three classes, i.e., hazardous, not hazardous, and moderate. In the first phase, the first hidden layer of our network, we used 50 neurons in the first hidden layer and 20 neurons in the second. In the second phase, we increased the neurons to 100 and 50 in both layers. We increased the number of neurons and not the number of layers to avoid overfitting, because the size of our data set is 2772 records of 180 days only. Moreover, after applying our proposed aggregation techniques to our data set, the records were reduced to 80-90, so we did not use a network that has many hidden layers.

We use “Cross-Entropy” as the loss function in our network. It is one of the commonly used loss functions for classification problems. Cross-entropy, also referred to as logarithmic loss or log loss, is a widely used loss function in machine learning to assess classification model performance. It quantifies the dissimilarity between the actual probability distribution and the model’s predicted probabilities. We present the results obtained from GCN in Table 3. The initial results produced by GCN do not show high accuracy. We understand that the small size of the data is a reason for these low values. During our data set preparation phase,

we learned that the longitude and latitude values coming from both types of data, i.e., cyber- and real-world facts, had very less common points, and many values were at different granularity levels of the same regions of South Korea.

We reconfigure our initial GCN architecture with two additional fully connected layers to aim for higher accuracy. It comprises of two convolutional layers, each with 200 neurons, followed by two fully connected layers, each with 150 neurons. Adding more convolutional layers led to overfitting, suggesting that this configuration is optimal for the current data set. Initially, using only the convolutional layers, we achieved 65% accuracy. However, after incorporating fully connected layers and fine-tuning the model, we improved the accuracy to 80%, as depicted through ROC in Figure 8.

4.5 Results Discussion

Table 2 illustrates a comparison among various classical machine learning models. Notably, NB yields the lowest accuracy, followed by DT and SVM. Conversely, GBDT, an ensemble method of DT, outperforms the rest, with K-Nearest Neighbors (KNN) as the second-best performer. GBDT's superiority over its base DT technique stems from aggregating results from individual DTs to calculate the outcome. GBDT is particularly well-suited for binary classification tasks like ours, where the goal is to predict a location as hazardous or not. Furthermore, while a single DT may exhibit lower bias, it tends to have higher variance. Ensemble methods like GBDT address this issue by effectively reducing variance. Both algorithms demonstrate comparable accuracy in comparing GBDT and KNN, with GBDT slightly edging out KNN. This can be attributed to several factors. GBDT is a sequential algorithm that builds new trees to correct errors from previous trees, allowing it to capture complex data relationships more effectively. KNN, being an instance-based algorithm, may overlook the global data structure and struggle with intricate decision boundaries. Additionally, GBDT often performs well with high-dimensional data, while KNN's performance can be sensitive to the choice of the parameter K.

Besides the classical ML algorithms, we also observed competitive performance from Graph Convolutional Networks (GCN), as illustrated in Figure 8 and Table 3. While not consistently superior, GCN achieved results comparable to or surpassing Gradient Boosting Decision Trees (GBDT) and K-Nearest Neighbors (KNN). Adding fully connected layers to the convolutional layers enhances GCN's performance. These fully connected layers aggregate and transform features captured by the convolutional layers, enabling the model to leverage the complete feature vector of each node (representing a region in our dataset) from across the entire graph. This allows the model to capture global dependencies and relationships that are not limited to local neighborhoods. In contrast, while effective at capturing local neighborhood information, the convolutional layers may struggle to integrate global information effectively, even with an extended receptive field.

In conclusion, our analysis reveals a noteworthy comparison between the performance of classical machine learning algorithms and Graph Convolutional Networks (GCN) on our curated dataset. While GCN may not have outperformed all traditional ML algorithms, its application and the release of our dataset pave the way for exciting new research avenues in this field.

5. Conclusion and Future Work

In this paper, we present a novel data-driven approach to establish a meaningful connection between cyber world data sourced from popular social media platforms and real world facts obtained from reputable institutions. Our methodology involves modeling the data from these distinct domains as a directed multi-attributed graph, allowing us to pinpoint areas affected by various phenomena. Through experimentation, we observed a robust correlation between the number of individuals infected with COVID-19 and those traveling between locations. Having meticulously curated this data set through extensive efforts, we have made it publicly available to facilitate further research by data scientists worldwide.

Encouraged by our initial findings, our future endeavors include expanding the dataset to incorporate information from additional social platforms, such as OpenStreetMap (OSM) traces. Furthermore, we plan to explore various graph modeling approaches, employing various combinations of graph-convolutional network (GCN) architectures to enhance the accuracy and effectiveness of our methodology.

Availability of data and materials

We provide the following links to publicly available data sets and processed data. The implementation source code will be available on request to interested researchers.

- **KCDC (Korea Centers for Disease Control and Prevention) Dataset:** `https://www.kdca.go.kr/index.es?sid=a3` `https://www.kaggle.com/kimjihoo/coronavirusdataset,2020`
- **META's Data for Good:** `https://dataforgood.facebook.com/dfg/tools`
- **Coronavirus (COVID-19) tweets dataset:** `https://dx.doi.org/10.21227/781w-ef42`
- **Processed dataset:** `https://github.com/WNawaz/Covid19HotspotPrediction`

References

[1] Z. Fu, Y. Wu, H. Zhang, Y. Hu, D. Zhao, and R. Yan, "Be aware of the hot zone: A warning system of hazard area prediction to intervene novel coronavirus covid-19 outbreak," Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.2241–2250, 2020.

[2] A. Kapoor, X. Ben, L. Liu, B. Perozzi, M. Barnes, M. Blais, and S. O'Banion, "Examining covid-19 forecasting using spatio-temporal graph neural networks," arXiv preprint arXiv:2007.03113, 2020.

[3] L. Wang, A. Adiga, S. Venkatramanan, J. Chen, B. Lewis, and M. Marathe, "Examining deep learning models with multiple data sources for covid-19 forecasting," 2020 IEEE international conference on big data (Big Data), pp.3846–3855, IEEE, 2020.

[4] F. Ucar and D. Korkmaz, "Covidiagnosis-net: Deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (covid-19) from x-ray images," Medical hypotheses, vol.140, p.109761, 2020.

[5] A.I. Khan, J.L. Shah, and M.M. Bhat, "Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images," Computer methods and programs in biomedicine, vol.196, p.105581, 2020.

[6] L. Qin, Q. Sun, Y. Wang, K.F. Wu, M. Chen, B.C. Shia, and S.Y. Wu, "Prediction of number of cases of 2019 novel coronavirus (covid-19) using social media search index," International journal of environmental research and public health, vol.17, no.7, p.2365, 2020.

[7] N. Darapaneni, D. Reddy, A.R. Paduri, P. Acharya, and H. Nithin, "Forecasting of covid-19 in india using arima model," 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp.0894–0899, IEEE, 2020.

[8] D. Liu, L. Clemente, C. Poirier, X. Ding, M. Chinazzi, J.T. Davis, A. Vespignani, and M. Santillana, "A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models," arXiv preprint arXiv:2004.04019, 2020.

[9] Z. Hu, Q. Ge, S. Li, L. Jin, and M. Xiong, "Artificial intelligence forecasting of covid-19 in china," arXiv preprint arXiv:2002.07112, 2020.

[10] M. Arashi, A. Bekker, M. Salehi, S. Millard, B. Erasmus, T. Cronje, and M. Golpaygani, "Spatial analysis and prediction of covid-19 spread in south africa after lockdown," arXiv preprint arXiv:2005.09596, 2020.

[11] N. Vindegaard and M.E. Benros, "Covid-19 pandemic and mental health consequences: Systematic review of the current evidence," Brain, behavior, and immunity, vol.89, pp.531–542, 2020.

[12] A. Kumar and K.R. Nayar, "Covid 19 and its mental health consequences," 2021.

[13] Y. Wang, L. Shi, J. Que, Q. Lu, L. Liu, Z. Lu, Y. Xu, J. Liu, Y. Sun, S. Meng, *et al.*, "The impact of quarantine on mental health status among general population in china during the covid-19 pandemic," Molecular psychiatry, vol.26, no.9, pp.4813–4822, 2021.

[14] P.S. Peixoto, D. Marcondes, C. Peixoto, L. Queiroz, R. Gouveia, A. Delgado, and S.M. Oliva, "Potential dissemination of epidemics based on brazilian mobile geolocation data. part i: Population dynamics and future spreading of infection in the states of sao paulo and rio de janeiro during the pandemic of covid-19," MedRxiv, pp.2020–04, 2020.

[15] M.R. Rahman, A.H. Islam, and M.N. Islam, "Geospatial modelling on the spread and dynamics of 154 day outbreak of the novel coronavirus (covid-19) pandemic in bangladesh towards vulnerability zoning and management approaches," Modeling earth systems and environment, vol.7, pp.2059–2087, 2021.

[16] M.A. Tewara, P.N. Mbah-Fongkimeh, A. Dayimu, F. Kang, and F. Xue, "Small-area spatial statistical analysis of malaria clusters and hotspots in cameroon; 2000–2015," BMC infectious diseases, vol.18, pp.1–15, 2018.

[17] A.M. Oster, G.J. Kang, A.E. Cha, V. Beresovsky, C.E. Rose, G. Rainisch, L. Porter, E.E. Valverde, E.B. Peterson, A.K. Driscoll, *et al.*, "Trends in number and distribution of covid-19 hotspot counties-united states, march 8–july 15, 2020," Morbidity and Mortality Weekly Report, vol.69, no.33, p.1127, 2020.

[18] Y. Zhang, X. Wang, Y. Li, and J. Ma, "Spatiotemporal analysis of influenza in china, 2005–2018," Scientific Reports, vol.9, no.1, p.19650, 2019.

[19] U. Qazi, M. Imran, and F. Ofli, "Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information," SIGSPATIAL Special, vol.12, no.1, pp.6–15, 2020.

[20] E. Chen, K. Lerman, E. Ferrara, *et al.*, "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set," JMIR public health and surveillance, vol.6, no.2, p.e19273, 2020.

[21] R. Lamsal, "Coronavirus (covid-19) tweets dataset," 2020.

[22] F.B. Hamzah, C. Lau, H. Nazri, D.V. Ligot, G. Lee, C.L. Tan, M. Shaib, U.H.B. Zaidon, A.B. Abdullah, M.H. Chung, *et al.*, "Coronatracker: worldwide covid-19 outbreak data analysis and prediction," Bull World Health Organ, vol.1, no.32, pp.1–32, 2020.

[23] T. Kuchler, D. Russel, and J. Stroebel, "The geographic spread of covid-19 correlates with structure of social networks as measured by facebook (2020)," tech. rep., CESifo Working Paper, 2020.

[24] "Korea data." `http://ncov.mohw.go.kr/`. Accessed: 2023-11-15.

[25] "Korea disease control and presentation agency." `https://www.kdca.go.kr/index.es?sid=a3`. Accessed: 2023-11-15.

[26] "National institute of health, korea." `https://nih.go.kr/eng/main/main.do`. Accessed: 2023-11-15.

[27] "World health organization." `https://www.who.int/`. Accessed: 2023-11-15.

[28] "Data science for covid-19." `https://www.kaggle.com/kimjihoo/coronavirusdataset,2020`. Accessed: 2023-11-15.

[29] "Republic of korea situation-who-covid." `https://covid19.who.int/region/wpro/country/kr`. Accessed: 2023-11-15.

[30] "Who coronavirus (covid-19) dashboard." `https://covid19.who.int/`. Accessed: 2023-11-15.

[31] "Korea-public health weekly report." `https://www.kdca.go.kr/board/board.es?mid=a30501000000&bid=0031&cg_code=C06`. Accessed: 2023-11-15.

[32] "Data for good." `https://dataforgood.facebook.com/`. Accessed: 2023-11-15.

[33] "Build location intelligence into your app." `https://location.foursquare.com/products/movement-sdk/`. Accessed: 2023-11-15.

[34] "Public gps traces." `https://www.openstreetmap.org/traces`. Accessed: 2023-11-15.

[35] "Us counties: Covid19 + weather + socio/health data." `http://tinyurl.com/4rajubc8`. Accessed: 2023-11-15.

[36] M. Ye, P. Yin, and W.C. Lee, "Location recommendation for location-based social networks," Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, pp.458–461, 2010.

[37] H. Wang, M. Terrovitis, and N. Mamoulis, "Location recommendation in location-based social networks using user check-in data," Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems, pp.374–383, 2013.

[38] "Meta-data for good." `https://dataforgood.facebook.com/dfg/tools`. Accessed: 2023-11-15.

[39] M.A. Lozano and F. Escolano, "Acm attributed graph clustering for learning classes of images," Graph Based Representations in Pattern Recognition: 4th IAPR International Workshop, GbRPR 2003 York, UK, June 30–July 2, 2003 Proceedings 4, pp.247–258, Springer, 2003.

[40] S. Abu-El-Haija, A. Kapoor, B. Perozzi, and J. Lee, "N-gcn: Multi-scale graph convolution for semi-supervised node classification," uncertainty in artificial intelligence, pp.841–851, PMLR, 2020.

[41] R. Zhang, Y. Li, A.L. Zhang, Y. Wang, and M.J. Molina, "Identifying airborne transmission as the dominant route for the spread of covid-19," Proceedings of the National Academy of Sciences, vol.117, no.26, pp.14857–14863, 2020.

Waqas Nawaz (Member, IEEE) received Ph.D. degree from Kyung Hee University, Republic of Korea, in 2015. He worked as a Post-Doctoral Fellow in the Department of Computer Science at Innopolis University Russia for one year. He is an Associate Professor with the Department of Computer and Information Systems, Islamic University of Madinah, Saudi Arabia. He actively participates in various research activities as a reviewer, guest editor, and researcher. His research interests include data mining, social networks analysis, big data, graph mining, databases, image processing, and artificial intelligence.

Muhammad Uzair received his master's degree in data science from FAST NUCES Islamabad, Pakistan, in 2021. He is serving as a Data Scientist in Teradata Corporation which is an American Software company that provides cloud database and analytics-related software. His research interests are Artificial Intelligence, Big Data and Data mining.

Kifayat Ullah Khan is serving as a Senior Lecturer in the College of Accounting, Finance and Economics, Birmingham City Business School, Birmingham City University, Birmingham, UK. He did his Ph.D. from Kyung Hee University, South Korea, and his MS from the University of Greenwich, UK. His research interests are Artificial Intelligence, FinTech, Databases, Big Data, and Social Network Analysis.

Iram Fatima received her Ph.D. degree from Kyung Hee University, South Korea, in 2014. Following her doctorate, she joined King Faisal University as an Assistant Professor, a position she held for five years. Even now, she remains actively involved in research activities at the university, contributing her expertise to various projects. Currently, she is also working as a Data Scientist in the IT industry in Canada. Her research interests are natural language processing, data mining, big data analytics, and artificial intelligence.