Spatio-Temporal Prediction in Epidemiology Using Graph Convolution Network

Dhanyalaxmi Panickar*, G P Sajeev†, S Siji Rani ‡
*‡Dept of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
†Dept of Electronics & Communications Engg., Govt Engineering College Wayanad, India
Email: *amenp2ari20012@am.students.amrita.edu, †gpsajeev@gecwyd.ac.in, ‡sijiranis@am.amrita.edu

***Abstract*—Infectious disease outbreaks are critical threats to public health and national security. With the significant increase in travel, trade, and human migration, contagious diseases can spread quickly over large areas and cause a severe epidemic. A timely and accurate prediction of pandemic outbreaks is crucial to detecting their outbreak, designing appropriate policies, averting further spread, and reducing losses. Pandemic data being highly nonlinear and complex, traditional methods cannot meet the requirements of medium and long-term prediction tasks and often neglect spatial and temporal dependencies. This work proposes a spatio-temporal epidemiological model for COVID-19 death prediction that uses Graph Convolution Network, demographic of the region, and mobility data. Our proposed method learns from a spatio-temporal graph where the nodes represent the region, spatial edges represent the distance and human mobility between the areas, and temporal edges represent node features through time. Our experiments are conducted on Johns Hopkins University (JHU) Coronavirus Resource Center data and are evaluated based on MSE. Our graph neural network model leverages spatial and temporal data to learn epidemiology's complex dynamics, helping it perform better than our considered baseline models- XGboost, Prophet, compartmental models like SRIMAX, and neural network model.**

***Index Terms*—Pandemic prediction, Deep learning, Graph convolution network, Spatio-temporal data**

I. Introduction

In recent decades many new infectious diseases, such as COVID-19, H1N1 influenza virus, severe acute respiratory syndrome(SARS), and Middle East respiratory syndrome coronavirus(MERS), have appeared and increased global public health crises [1]. With no existent immunity in the population and greater human mobility, these new infectious diseases spread more quickly and efficiently, leading to substantial severe cases and mortality in the community. Such infectious disease outbreaks severely impact a nation's security and economic conditions. Especially for a highly communicable new respiratory infectious disease, medical help, including drugs, personal protective gear, and life support, can quickly deplete once hospitals are overwhelmed with infected patients. It can inevitably cause excess mortality, as demonstrated in numerous countries during the COVID-19 pandemic. Figure 1, shows us the history of infectious disease outbreaks and their corresponding death tolls.

However, in the early stage of a novel infectious disease outbreak, there is usually no prior knowledge and available procedures to guide the governing bodies in making decisions. Here we could exploit machine learning technologies to tackle the issue. Recent studies have shown that the application of AI is a game-changer due to its rapid processing capacity and enhanced efficiency [2]. Predictive analytics of infectious disease permits us to estimate when and under which circumstances countries can expect increases, peaks, and reductions in new cases and deaths. With this data, the governing agencies can be prepared for the surge in demand for medical services and decide the duration and intensity of containment measures.

Epidemiological models can be broadly classified as Compartment, Metapopulation, Agent-Based, and Network models [3] [4]. Compartment models divide the population and assign it to compartments with different labels, e.g., S - Susceptible, I - Infectious, R - Recovered. The order of the labels signifies the flow pattern between the compartments, e.g., SI, SIR, SIRS. Metapopulation models introduce an addition to the spatial structure of the population. Agent-Based models [5] are computer simulations composed of agents that can interact with each other and an environment. Few parameters determine compartmental and agent-based models, making them inefficient in capturing complex infectious disease patterns of the infectious disease. The other models studied built separate models for each location, where the forecast is based only on the data from that location without considering the geographical proximity of other locations and mobility. The future number of infection cases and deaths of the infectious disease depends on its historical information and other areas', people traveling to/out of this area, and areas with similar epidemic patterns, to name a few. Hence we can consider enhancing the forecast accuracy by using real-time data representing inter-region mobility and by developing a unifying approach that can encompass both the temporal and spatial interactions for infectious disease modeling. In this work, we propose a model that aims to predict the number of deaths due to infectious diseases utilizing the spatiotemporal feature of the data. We map locations to nodes on a graph, construct an edge, and provide edge weight based on the geographical proximity and number of flights between locations. Each node has static and dynamic features based on the past death count and the location's demographic. Then we use GCN to utilize the spatio-temporal graph to forecast the deaths due to COVID-19 in one month's future.

The organization of the rest of the paper is as follows. Section II reviews some of the standard approaches, their

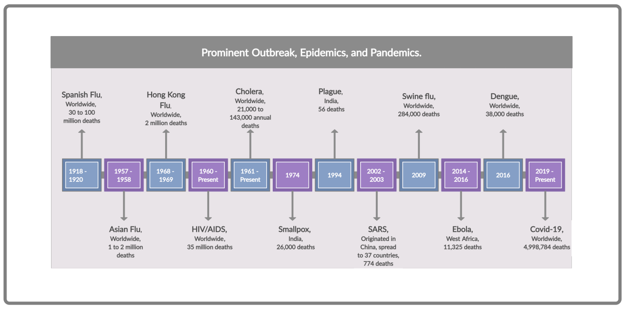
Prominent Outbreak, Epidemics, and Pandemics.
Spanish Flu, Worldwide, 30 to 100 million deaths
Hong Kong Flu, Worldwide, 2 million deaths
Cholera, Worldwide, 21,000 to 143,000 annual deaths
Plague, India, 56 deaths
Swine flu, Worldwide, 284,000 deaths
Dengue, Worldwide, 38,000 deaths
1918 - 1920
1957 - 1958
1968 - 1969
1960 - Present
1961 - Present
1974
1994
2002 - 2003
2009
2014 - 2016
2016
2019 - Present
Asian Flu, Worldwide, 1 to 2 million deaths
HIV/AIDS, Worldwide, 35 million deaths
Smallpox, India, 26,000 deaths
SARS, Originated in China, spread to 37 countries, 774 deaths
Ebola, West Africa, 11,325 deaths
Covid-19, Worldwide, 4,998,784 deaths

Fig. 1. Infectious diseases over the years

merits, and demerits while also establishing gaps and motivations for this work. Section III presents the methodology and implementation steps. Section IV is subdivided into three subsections - Dataset, Metrics used, and Results. Finally, section V presents the conclusion and future works.

II. Related Work

Accurate forecasts for the pandemic outbreak and its effect could help reduce its impact and allow governments to refine their prevention strategies and advance the necessary safety measures. Hence a relatively new area of Public Health, digital epidemiology, has acquired recognition, providing effective monitoring of confirmed cases, overcrowding, and over-death. Many studies have been done for epidemiological forecasting using compartmental, statistical, machine learning, and hybrid approaches [6].

This work began with the classical paper of Kermack–McKendrick in 1927. This work laid the foundation for compartmental models. Kermack–McKendrick's theory became the source for SIR models and their related models [7]. Suspected-infected-recovered (SIR)-type classic epidemiological models mainly facilitate forecasting the number of cases over a specified time interval. Due to the significant spatial differentiation of region forecasting, the number of cases limited to a scalar numeric value on a given day for the entire country is wholly insufficient. Many models have been developed by making advancements in the classical compartmental model [1]. However, these models are still determined by only a few parameters(e.g., reproduction number, transmission rate, recovery rate), making them unable to capture complex patterns [8]. In some studies, separate models were built for each location without including geographic proximity and mobility between regions. Alternatively, the predictions were based on some patterns observed in other areas. [9]. A location often shows similar disease patterns with its nearby locations or demographically similar locations due to population movements or demographics.

Based on the review, we infer that forecast accuracy can improve by utilizing the inter-region interactions and mobility data and developing a unifying technique that can enclose both the temporal and spatial interactions for infectious disease modeling.

III. Proposed Model

More accurate predictions of infectious diseases can be made using the spatial and temporal features of the data. As graphs are realistic representations of a wide variety of real-life data in social, biological, financial, and many other fields, we seek to utilize the quality to model our deep learning model. We aim to build and train a model that can understand spatial data's importance and utilize the temporal data to forecast the deaths due to COVID-19 in one month's future.

In the proposed work, we are using the Graph Convolution Network (GCN) to understand the features and importance of the spatial data and how it affects the spread of COVID-19 over the given period. For the GCN to work and predict the future, we must provide the graph data. However, the considered data are all in the tabular format. So we have to build a pipeline to convert the tabular data into graph data using the PyTorch Geometric Temporal module.

We can divide our project into three phases: Data pre-processing, processed data to graph conversion, and GCN implementation.

A. Data pre-processing

Figure 2, shown below, is our proposed model's workflow for data processing. We are considering multiple datasets to create our graph dataset. Here, the first dataset (COVID-19 statistics data provided by Johns Hopkins University Coronavirus Resource Center) is about the covid-19 international spread and deaths data. The second dataset (U.S. population data (2019) by The U.S. Census Bureau) is for understanding the State-wise population and senior citizen population of the U.S. Country. Lastly, the third dataset (U.S. domestic flight data from the Bureau of Transportation Statistics) is for understanding the recorded State-wise flight travels. Data acquisition means loading/importing the necessary data into the python workspace and converting the standard tabular data like CSV files into understandable data such as the" ndarray" object of the" NumPy" object. After data acquisition, we perform data analysis to understand the basics, such as the number of rows and columns and the type of data in each column of loaded data, so that we can perform the data pre-processing step efficiently. In the data pre-processing step, we clean and prepare the data for giving it as input to the algorithm. For pre-processing the raw international COVID-19 dataset, we perform feature selection, grouping, and windowing on it.

1) Feature Selection: : We will select the columns and rows per the desired task in the tabular data. Here we are selecting all the necessary columns to get the data about all the states in the U.S.A.

2) Grouping: : Understanding the groups of data with respect to a particular condition. Here we are grouping the data for a considered period (Jan2020 to Dec 2020) for every single state in the U.S.A.

3) Windowing: : For a given period, understanding the values of each row transposed into a single row becomes a new set of features that can give us an understanding of how values change over a given time.

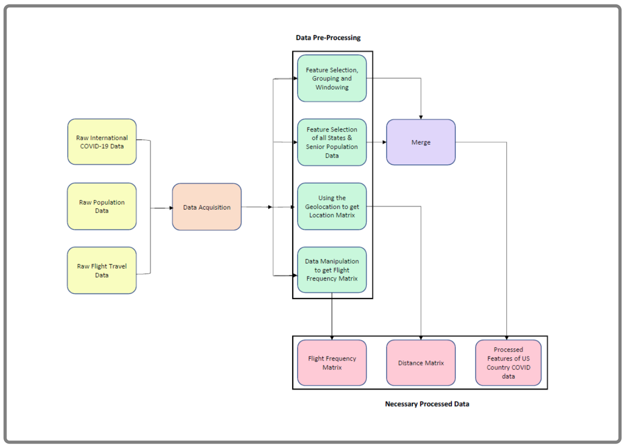
Data Pre-Processing
Raw International COVID-19 Data
Raw Population Data
Raw Flight Travel Data
Data Acquisition
Feature Selection, Grouping and Windowing
Feature Selection of all States & Senior Population Data
Using the Geolocation to get Location Matrix
Data Manipulation to get Flight Frequency Matrix
Merge
Flight Frequency Matrix
Distance Matrix
Processed Features of US Country COVID data
Necessary Processed Data

Fig. 2. Data Pre-Processing Workflow

4) Data Manipulation: : For any given data, if the data structure is completely changed to a new, more helpful structure, then it is called data manipulation. For our project, we are converting the raw Covid-19 standard tabular data into a square matrix for the distance between two states. The distance between the two states is calculated based on the longitude and latitude of the two states. The raw flight travel standard tabular data is converted into a square matrix for the flight frequency between two states.

The features selected from the Covid-19 data and Population data are merged to get the proceeded features of U.S. country covid data. The features selected for our graph nodes are 12 months of death cases, population count, density, and percentage of population above the age of 65. While there will not be a group of the population untouched by this crisis, the elderly population is likely to face the worst effects. Initial Provisional Death Counts for Coronavirus Disease 2019 reports of C.D.C. have shown that 80% of the deaths due to COVID-19 occur in those over the age of 65. Since the virus has largely affected the elderly, we have added the percentage of the population above the age of 65 as a node feature for our graph.

B. Processed Data to Graph Conversion

Figure 3 illustrates the workflow for the processed data to graph conversion workflow. To convert any data to a graph, we first have to identify these entities- nodes, node features, edges, and edge weights(optional). The first step to generating graph data is getting the base location data for a selected country and understanding the sub-locations of the country to form the nodes of the graph.

In our proposed model, 50 states of the U.S. country are considered nodes. The past 11 months of the data for each state are considered the primary features of the nodes. Population Count, Population Density, and Percentage of people above the age of 65 are three other important features (normalized) considered to weigh the node's importance. The node features are divided into X and Y features, where the Y feature represents the number of death cases in the 12th month, and the remaining features represent the X features.

Furthermore, we are considering a fully connected graph structure. Each node will be connected to every other node, creating n*(n-1)/2 edges for n is the number of nodes. Hence each node (states) will have an edge connecting each other. Every edge in our graph is weighted based on two parameters: Inverse distance and Flight count between the nodes of the edges. We compute the distance by calculating the geodesic distance between the states using the latitude and longitude coordinates of the states. Hence, closer nodes (states) will have a higher inverse distance value, and nodes with more flights going to and fro between them will have a higher value. This way, we get large edge weight values for related nodes (states). As input to data to graph conversion block, we will be passing states as graph nodes and the respective number of edges, Which will be fed to the GCN, which in turn will help us to predict the future values(deaths due to Covid-19) of nodes(state) for the 12th month. Data Splitting means creating the training set and testing set so that we can train the algorithm and understand the performance of that model. In our proposed project, we are taking an 80:20 ratio of data for training and testing, respectively. We split the data using node masking. Now we have all the components we need to build a graph for libraries like PyTorch Geometric Temporal. We will pass the NumPy arrays for node, node features, edges, and edge features to the Data object. This data object represents one single graph. This data object is then passed to the PyTorch Geometric Temporal data loader to create an Inmemory graph training and test dataset graph dataset.

C. Graph Convolution Network

Graph machine learning has the primary challenge of finding a way to represent graph structure so that it can be easily exploited by machine learning models; in particular, the Graph Convolutional Networks (GCN) are the generalization of the standard Convolutional Neural Networks, used to extract low-level features from image data, and are based on the idea that not only does the best nodes' representation depend by their own characteristics, but also by their neighbourhood description and topology.

Formally, let G=(V,E) be a generic graph, V,E being the nodes and edges sets respectively, and consider its adjacency matrix A; then each layer H(l) is defined recursively as follows [10]:

$$H^{(l+1)} = f(H^{(l)}, A) = \sigma(\hat{D}^{-1/2} \cdot \hat{A} \cdot \hat{D}^{-1/2} \cdot H^{(l)} \cdot W^{(l)}), \quad (1)$$

Sigma being an activation function, D being the nodes diagonal node degree matrix of A, W being the (learnable) weight matrix for the l-th layer.

The spatio-temporal GCN used in our model is composed of several spatio-temporal convolutional blocks, each of which is formed as a "sandwich" structure with two gated sequential convolution layers and one spatial graph convolution layer in between.

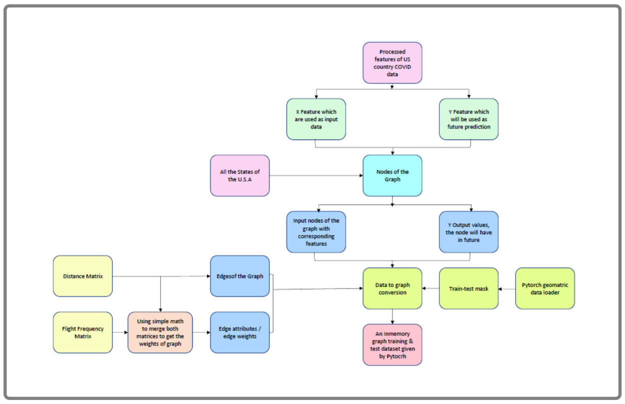
Processed features of US country COVID data
X Feature which are used as input data
Y Feature which will be used as future prediction
All the States of the U.S.A
Nodes of the Graph
Input nodes of the graph with corresponding features
Y Output values, the node will have in future
Distance Matrix
Edgesof the Graph
Data to graph conversion
Train-test mask
Pytorch geomatric data loader
Flight Frequency Matrix
Using simple math to merge both matrices to get the weights of graph
Edge attributes / edge weights
An Inmemory graph training & test dataset given by Pytocrh

Fig. 3. Processed Data to Graph Conversion Workflow

IV. Performance Evaluation

The proposed model is evaluated through simulations; for that purpose, a model has been developed in Python. Section IV-A elaborates on the Datasets used. Section IV-B explains the Metrics used for the evaluation of performance. Lastly, IV-C presents results and observations.

A. Dataset

The Covid-19 statistics data used in this project is provided by Johns Hopkins University (JHU) Coronavirus Resource Center. The covid-19 data from JHU Coronavirus Resource Center were collected from Jan 3, 2020, to Dec 3, 2020. U.S Flight travel data is obtained from the bureau of transportation statistics, which contains details about all domestic flights, such as source destination location and flight date for the year 2020.

B. Metrics Used

The performance of the proposed approach is evaluated based on three performance measures: MAE and RMSE.

1) Mean Absolute Error (MAE): The mean absolute error (MAE) represents the difference between the actual and the measured value. By evaluating this metric, we can get an idea of the precision of a value. Indeed, if we know the measured and actual values, we can perform a simple subtraction to find the absolute error. This is obtained using eq.(1)

$$MAE = \frac{1}{N}\sum_{t=1}^{n}|F_t - O_t| \tag{1}$$

C. Root Mean Squared Error (RMSE)

This root mean squared error (RMSE) is a statistical measure, which calculates the average magnitude of errors. It does not indicate their direction. In fact, the mean square error gives more weight to large errors than to others when calculating the mean, compared to the mean absolute error. So, when the root mean square error is much greater than the mean absolute error, it means an increased frequency of large errors. It is therefore an appropriate statistical when large errors are very undesirable. The RMSE is calculated using eq.(2).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(F_t - O_t)^2}{N}} \tag{2}$$

D. Results

The performance of the proposed model is compared against the four baseline models- SARIMAX, Prophet, XGBoost, and Neural Network.

Table I records the performance of our proposed model along with four baseline models used.

TABLE I
Evaluation metrics results

Model	RMSE	MAE
SARIMAX	4.52	4.18
Prophet	5.68	4.15
XGBoost	8.54	6.23
NN	6.12	5.25
GCN	3.87	3.38

It is observed that spatio-temporal GCN model performs better than the statistical and neural network models.

V. Conclusion

This work proposes a Covid-19 death prediction model. We have employed a graph neural network based approach for forcasting with spatio-temporal data. Also, our model doesn't rely on the assumption of the underlying disease dynamics and learns from interregional interactions. From the results, it could be ascertained that our GCNbased model has performed better than the statistical (SARIMAX) and neural network (ANN) models.

For future research, we could develop these results by incorporating new features, expanding the time horizon for long-term predictions, and experimenting on epidemiological mobility data in other parts of the world.

References

[1] K. Sarkar, S. Khajanchi, and J. J. Nieto, "Modeling and forecasting the covid-19 pandemic in india," *Chaos, Solitons Fractals*, vol. 139, p. 110049, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096007792030446X

[2] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, Jun. 2019. [Online]. Available: https://doi.org/10.7861/futurehosp.6-2-94

[3] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U. Reuter, T. Rabczuk, and P. M. Atkinson, "Covid-19 outbreak prediction with machine learning," *Algorithms*, vol. 13, no. 10, p. 249, Oct 2020. [Online]. Available: http://dx.doi.org/10.3390/a13100249

[4] A. K. Gupta, V. Singh, P. Mathur, and C. M. Travieso-Gonzalez, "Prediction of covid-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in indian scenario," *Journal of Interdisciplinary Mathematics*, vol. 24, no. 1, pp. 89–108, 2021. [Online]. Available: https://doi.org/10.1080/09720502.2020.1833458

[5] A. I. Saba and A. H. Elsheikh, "Forecasting the prevalence of COVID-19 outbreak in egypt using nonlinear autoregressive artificial neural networks," *Process Saf. Environ. Prot.*, vol. 141, pp. 1–8, Sep. 2020.

[6] S. D. Khan, L. Alarabi, and S. Basalamah, “Toward smart lockdown: A novel approach for COVID-19 hotspots prediction using a deep hybrid neural network,” *Computers*, vol. 9, no. 4, p. 99, Dec. 2020. [Online]. Available: https://doi.org/10.3390/computers9040099

[7] O. V. Volodina and https://pnojournal.wordpress.com/2022/07/01/volodina-3/, “Formation of future teachers’ worldview culture by means of foreign-language education,” *P Sci Edu*, vol. 57, no. 3, pp. 126–159, Jul. 2022.

[8] ——, “Formation of future teachers’ worldview culture by means of foreign-language education,” *P Sci Edu*, vol. 57, no. 3, pp. 126–159, Jul. 2022.

[9] S. Pei and J. Shaman, “Initial simulation of SARS-CoV2 spread and intervention effects in the continental US,” Mar. 2020.

[10] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016. [Online]. Available: https://arxiv.org/abs/1609.02907