# Machine Learning Approaches for Region-level Prescription Demand Forecasting

Wang, X., Chen, L., Wang, J. & Zheng, D



DOI 10.1109/SWC57546.2023.10449058
ISBN 9798350319804

Publisher: Institute of Electrical and Electronics Engineers Inc.

# Machine Learning Approaches for Region-level Prescription Demand Forecasting

1st Xu Wang
*Centre for Intelligent Healthcare*
*Coventry University*
Coventry
wangx197@uni.coventry.ac.uk

2nd Long Chen
*School of Computing, Mathematics and Data Sciences*
*Coventry University*
Coventry
ad8579@coventry.ac.uk

3rd Jiangtao Wang*
*Centre for Intelligent Healthcare*
*Coventry University*
Coventry
ad5187@coventry.ac.uk

4th Dingchang Zheng
*Centre for Intelligent Healthcare*
*Coventry University*
Coventry
ad4291@coventry.ac.uk

***Abstract*—Region-level prescription demand is closely intertwined with the incidence of diseases within a given area. However, conventional forecasting methods primarily rely on historical data, and ignore the spatial correlation in prescription data. In this study, we employ graph structures to capture the interactions among drug demand in different regions. By leveraging two popular graph neural network-based models, our objective is to harness the power of spatial-temporal correlation to enhance the accuracy of predictions. To assess the effectiveness of the graph neural network-based model, we conduct extensive experiments on a comprehensive real world dataset. The results demonstrate that the performance of the graph neural network consistently surpasses that of statistical learning-based methods and traditional deep learning-based methods.**

***Index Terms*—Region-level prescription demand, spatial temporal correlation, graph neural network**

## I. Introduction

In today's fast-paced world, the integration of technology into various aspects of our lives has become increasingly prevalent. One such area is healthcare, where technology plays a vital role in improving patient outcomes, optimizing resource allocation, and enhancing decision-making processes. Region-level prescription forecasting, in conjunction with Ubiquitous Intelligence and Computing (UIC), has emerged as a significant development, revolutionizing the way healthcare systems operate.

The prescription demand in a region is closely correlated with the prevalence of diseases. An increase in the demand for drugs signifies a rise in the number of local patients seeking treatment. Anticipating prescription demand in advance can serve multiple crucial purposes. Firstly, it enables the prevention of the further spread and deterioration of diseases by ensuring a timely and adequate supply of medications. Secondly, early prediction of prescription demand facilitates the effective allocation of limited medical resources by allowing the government to plan and allocate resources in a more efficient manner.

This work is supported by EPSRC New Investigator Award (Grant Ref: EP/V043544/1).

While humans have recorded drug consumption data for an extensive period, the amount used for each prescription was systematically recorded and categorized until recent years. However, most prescription data currently available also presents many problems when used in prediction models. One of the biggest challenges of using prescription demand data is the frequent occurrence of incorrect or missing data in the dataset, which will inevitably lead to errors in the model output. Furthermore, prescription data can be influenced by a variety of factors that are often interrelated in consecutive time and locations. Focusing solely on a single region or timeslot hampers the accurate simulation of actual prescription demand. To overcome these challenges, it is crucial to account for the spatial and temporal correlation of the data in prediction. By considering these correlations, we can obtain more precise and reliable forecasts.

The current approach to forecasting prescription demand relies heavily on expert opinion. However, hiring consultants can be costly, and relying solely on expert opinion can be less objective due to biases and personal experiences. There are various predictive models available, many of which have demonstrated their effectiveness in different domains. However, few studies have applied these models to prescription demand analysis. At present, predictive models can be classified into distinct groups based on their methods, all of which have demonstrated their effectiveness in their respective fields, with graph neural networks being

particularly valuable in prescription demand prediction. Graph neural networks have been shown to be more effective in extracting knowledge from graph-structured data that is challenging for traditional neural networks. They are effective in extracting latent spatio-temporal correlations in the data, which aligns well with the needs of prescription drug forecasting.

This paper aims to analyze regional populations health status and apply regional prescription demand forecasting. The primary objective is to assess the suitability of different forecasting models for prescription demand analysis. Specifically, the study explores the spatio-temporal correlation of prescription data and examines the feasibility of multiple prediction models. The main contributions of this article are as follows:

- Identify the problem of regional prescription demand forecasting and modeling the spatial correlation between prescription data in different locations using a graph structure.
- Analyze the feasibility of models including statistical learning models, deep learning models and graph neural network models in the context of prescription demand forecasting.
- We compare the performance of these models on a real-world dataset, which shows that ASTGCN can consistently outperform the other models.

## II. Related Works

### A. Medication demand analysis

Scholars have shown considerable interest in studying the trend of prescription demand. Björn et al. [1] developed a linear regression model based on historical data to forecast the growth rate of drug spending in Stockholm, Sweden. However, regional drug consumption is often influenced by multiple factors, and it is challenging for linear models to precisely predict changes in drug consumption over time.

In order to improve prediction accuracy, scholars began to introduce different statistical methods to process the prescription demand data. In [2], trends of prescription demand data have been analyzed using a variety of statistical methods. It classifies and screens out drugs with a large amount of use to capture the pattern of prescription consumption. On the prescription consumption data collected in household interviews, they used survey-weighted logistic regression and join-point analyzes to reveal the average annual percentage change of various types of drugs among American adults.

As machine learning continue to evolve, AI models have started to be used in the analysis of prescription demand data. Bora et al. [3] combined the epidemiological model and bidirectional LSTM to estimate the demand for medicines required by the hospital. Compared with the traditional ridge regression model, the bidirectional LSTM model shows better performance. The paper [14] proposes an illicit drug prediction method based on sales data from darknet markets. On the basis of historical data, it adds Wikipedia's drug page views data, which effectively improves the prediction accuracy. However, this approach is only effective for new drugs.

### B. Spatial-temporal analysis for population health

Population health data in a given area exhibits a strong association with the corresponding prescription consumption. Although the literature on the analysis of prescription demand in specific regions is relatively limited, previous studies have explored the spatial-temporal correlations of population health data to identify underlying patterns. In particular, several studies focused on understanding the spatial-temporal variation of diseases, providing valuable insights into factors that may affect prescription demand in different regions.

In many instances, population health data for neighboring regions and time steps are highly correlated. In order to confirm and utilize the spatiotemporal correlation of disease incidence data, the paper [4] proposes a compressive population health method, which can infer the incidence of chronic diseases in adjacent areas by using a small amount of data. Convolutional neural networks and generative adversarial networks are employed in this method to model the spatiotemporal correlations of incidence, thereby greatly reducing the cost of data collection. Using data on confirmed COVID-19 cases, deaths, and population movements across the United States, the authors [5] propose a multivariate predictive model. Due to the combination of spatiotemporal correlations of diseases on traditional population migration data, it has a significant improvement in accuracy compared to other prediction methods. Similarly, the paper [6] employs graph neural networks to model the spatial-temporal correlation contained in the COVID-19 infection data. With graph neural network, the relationships between data can be further explored.

Considering the hysteresis of disease data statistics, the paper [7] models both real-time data and updated data separately, and constructed a POPNET system. This method utilizes recurrent neural network and graph attention mechanism to extract the spatial and temporal correlation of population health data. By adaptively updating the dataset to optimize the model, it results in less prediction error in real datasets.

## III. Dataset and Problem Formulation

### A. Dataset

The dataset we use in this paper is the England Practice level prescribing data published by the NHS [1] . The number

[1]Practice Level Prescribing Data available at: https://digital.nhs.uk/data-and-information/publications/statistical/practice-level-prescribing-data

of prescriptions filled in each practice from August 2010 to December 2019 was collected in this dataset. Specifically, this dataset contains the addresses and codes of practices, number of prescriptions, prescription names and their corresponding ICF-10 codes.

During the course of a decade, some practices have undergone relocations, resulting in incomplete data availability for certain periods. Additionally, the demand for some prescriptions is close to 0 most of the time, which is not conducive to prediction. However, for drugs that treat chronic diseases, such as diabetes, the demand for prescriptions exists throughout the time period. To ensure the validity of the predictions, we extracted practices that did not change geographically over the past ten years and selected specific drugs for which there was a consistent demand for experiments. After the dataset was processed, we conduct the experiments with three drugs on a dataset of 513 practices in London.

In the experiment, we divided the prescription demand data into three parts for training, validation and testing. The training set contains 73 months of prescription statistics from August 2010 to August 2016. The validation set contains prescription statistics for 20 months from October 2016 to month April 2018. The test set contains prescription statistics for 20 months from May 2018 to December 2019.

### B. Problem Formulation

*1) Prescription network G:* We formulate the prescription demand prediction problem as an undirected graph $G = (V, E, A)$, where $V$ represents a set of $N$ nodes and $E$ is a set of edges. $A$ is the adjacency matrix of the graph $G$, which is generated from the connectivity between nodes. If two nodes $i$ and $j$ are connected, $A_{i,j}$ is 1. In this paper, we define a threshold $\varepsilon$ to compute the adjacency matrix $A$:

$$A_{i,j} = \begin{cases} 1, & if\ exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) \leq \omega \\ 0, & otherwise \end{cases} \tag{1}$$

where $d_{ij}$ is the euclidean distance between node $i$ and node $j$, and $\sigma^2$ is the hyperparameter to control the sparsity of the adjacency matrix $A$. If $\omega$ is fixed, a smaller $\sigma^2$ results in a sparser $A$, and thus a simpler graph to construct. However, when $\sigma^2$ is too small, the correlation between nodes cannot be accurately modeled. In this paper, we set $\sigma^2$ and $\omega$ to be 0.1 and 80, respectively.

*2) Prescription statistics data X:* We represent the collected prescription data as a tensor $X \in \mathbb{R}^{N \times T \times K}$ , where $N$ is the total number of nodes, $T$ denotes the total number of timeslots, and $K$ is the number of prescription types. We define the prescription demand of each node location $n$ at timeslot $t$ as a matrix $X_t = \left(x_t^1, x_t^2, \cdots, x_t^N\right) \in \mathbb{R}^{N \times K}$, and the prescription demand of each timeslot $t$ at each location $n$ as a matrix $X_n = \left(x_n^1, x_n^2, \cdots, x_n^T\right) \in \mathbb{R}^{T \times K}$.

*3) Area-Level Prescription Demand Prediction:* Given the prescription statistics $X$ on the prescription network $G$, our task is to construct a multi-task prediction framework $F$, so as to predict the demand of each node n for different prescriptions at the future $h$ steps of timeslots.

## IV. Models for Spatial Temporal Data analytic

### A. Approach overview

There are many models available for data processing, which have their own advantages in dealing with different kinds of data. In this paper, we have selected some representative prediction models and classified them into three categories: statistical models, sequential machine learning models and graphical neural network models.

### B. Statistical model

*1) Historical Average:* The historical average model is a simple but useful model for time series forecasting, which predicts the data trend by averaging the data over a time period in the past. Due to its simplicity and high efficiency, it is currently used as a baseline for evaluating the accuracy and effectiveness of more complex forecasting models in time series in various research fields.

*2) ARIMA:* The Autoregressive Integrated Moving Average (ARIMA) model was introduced by BOX and Jenkins in 1970 [8]. The basic idea is to transform the original time series into a stationary time series by taking the difference. The differenced series are then fitted to make predictions for future values. Because of its ability to model non-stationary time series and to deal with variations in the feature, it has become one of the most classic models in time series forecasting models.

However, the performance of ARIMA is often unsatisfactory for non-linear time series. Additionally, the model's tolerance to outliers is weak. Therefore, for accurate prescription demand prediction, more advanced and effective methods are needed.

### C. Sequential machine learning models

*1) XGBoost:* XGBoost is an advanced version of the Gradient Boosting Decision Tree (GBDT) model, originally proposed by Tianqi Chen [9]. Compared with GBDT, XGBoost can train models more efficiently, while achieving high accuracy and speed in various regression and classification tasks.

Unlike other models, the exceptional performance of XGBoost is attributed to the ensemble training of multiple decision trees. It uses the gradient boosting framework to iteratively train each decision tree sub-model and combine their predictions into a stronger overall model. The new decision tree will be trained using the residuals from the previous iteration, which is one of the keys to its speed

and efficiency. The objective function for defining the t-th decision tree is as follows:

$$L^{(t)} = \sum_{i=1} l(y_i, (\hat{y}_i^{(t-1)} + f_t(x_i))) + \Omega(f_t) \tag{2}$$

where $l(y_i, \hat{y}_i^{(t-1)})$ represents the loss function related to $y_i$ and $\hat{y}_i^{(t-1)}$, $\hat{y}_i^{(t-1)}$ represents the predicted value of sample $i$ by the first $t-1$ decision trees, $y_i$ represents the actual value of sample $i$, $f_t(x_i)$ represents the predicted value of sample $i$ by the $t-1$ decision tree, and $\Omega(f_t)$ represents the model complexity of the $t$-th tree.

It is worth noting that all decision trees in XGBoost are trained in parallel to increase training speed. Furthermore, in order to reduce model complexity and improve generalization, XGBoost also uses pruning techniques to remove unnecessary branches from the decision tree.

*2) GRU:* In most of the network, the temporal correlation is extracted through Recurrent Neural Network (RNN)-based models-based models. As one of the most well-known structures in RNN-based models, Gate Recurrent Unit (GRU) is widely used with its simple structure and well performance.

One of the essential features of a GRU is its gating mechanism., which allows a model to selectively update and reset its hidden state based on the input and previous hidden states. This selective gating mechanism is critical for the model's ability to learn long-term dependencies in the input sequence and helps to overcome the vanishing gradient problem that traditional RNNs suffer from.

The GRU model consists of three gates that control information flow, including update gate, reset gate and output gate. The update gate is used to control the number of hidden states updated at the current time step, the reset gate determines the number of hidden states to be forgotten, and the output gate controls the number of current hidden states used for output at the current time step. Through the combined use of update gates and reset gates, the GRU model can effectively capture the long-term dependencies in the input sequence, thereby improving the model performance.

### *D. Graph neural network models*

*1) GCN:* Graph Convolutional Network (GCN) is a deep learning model based on graph structured data. It extends the concept of convolutional operations to graphs, enabling the extraction of features from neighboring nodes associated with a given node. By continuously cascading the output of each layer, GCN can perform higher-level feature extraction. Compared to traditional convolutional neural networks, GCN can extract more information from graph data and thus better handle the relationships and interactions between nodes. The information propagation method between layers of GCN is:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{3}$$

where $\tilde{A} = A + I$, $\tilde{D}$ is the degree matrix of $\tilde{A}$, $H$ represents the feature of each layer, and $\sigma$ is the nonlinear activation function.

GCN is a flexible and powerful deep learning model that can effectively handle graphically structured data. It demonstrates powerful modeling capabilities, making it one of the most popular options for spatiotemporal data analysis. In this paper, we have selected two of the most representative GCN model frameworks for spatiotemporal data analysis applications to study.

*2) STGCN:* STGCN was initially used to solve the traffic flow prediction problem. As previous models modeled the data by gridding it, usually only local features could be extracted, ignoring the connectivity and global nature of the whole traffic network. Yu et al. [10] proposed the STGCN model to extract higher-order features from graph-structured data in space. The main advantage of STGCN is its ability to process spatial-temporal features from the data in a more holistic and comprehensive way.

STGCN consists of two main components: spatial domain convolution and temporal domain convolution. Spatial domain convolution is mainly used to process spatial information in graph structures, while temporal domain convolution is responsible for processing time series data. By combining these two components, STGCN is able to perform high-order feature extraction on graph-structured spatial-temporal data.

*3) ASTGCN:* The traditional GCN can only handle graphs with fixed topology, and considers the information of the node itself and its immediate neighbor nodes when updating the node features. ASTGCN [11] improves the expressiveness and prediction performance of the model by introducing gating and attention mechanisms, while handling graphs with variable topology more effectively.

ASTGCN divides the input time series into three different time intervals, which are fed into three components with the same network structure for prediction. Each of these components contains multiple temporal convolution modules. In the time-domain convolution, ASTGCN also introduces a gating mechanism to control the flow of information, allowing the model to better capture long-term dependencies in the spatio-temporal data. The final prediction is obtained by combining the outputs of the three components by a parameter matrix.

## V. Results and Analysis

*1) Performance Metrics:* To comprehensively evaluate the performance of each prediction model, we employ two metrics: mean absolute error (MAE) and root mean

squared error (RMSE) [12], [13]. These metrics allow us to objectively assess the predictive accuracy of the model. They are calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{5}$$

*2) Data Preprocess:* In order to simplify the problem, we only analyze the prescription data in the London area in the experiment. There are 513 GP clinics in the London dataset for a total of 113 monthly prescriptions demand data. Our data preprocessing pipeline involves outlier detection and subsequent missing values reconstruction.

To begin, we employ the Z-score outlier detection method to identify outliers within the dataset. Any detected outliers are considered as missing values and subsequently treated accordingly. By setting these outliers as missing values, we ensure they do not influence our analysis. Next, we utilize the temporal average method to impute the missing values in our dataset. It is worth noting that for locations with missing values, we retain their data as input to the model to provide as much information as possible for the prediction, but do not perform an error assessment on their predicted values. Finally, as part of our evaluation process, we assess the predicted values of future prescription demand for all rest 379 GP clinics.

*3) Performance on Prescription Demand Prediction:* In this section, we investigated the effectiveness of different models in predicting prescription demand. We chose two chronic diseases, hyperlipidemia and diabetes, and conducted experiments on a drug corresponding to each of them. The results of each model are shown in table I and II.

Among all the models, the two graph neural network-based models, STGCN and ASTGCN, demonstrated superior performance across all evaluation metrics. For the prediction of hyperlipidemia drug, ASTGCN achieved a reduction of 21.4% and 20% in MAE compared to the optimal statistical model HA and the optimal deep learning model GRU, respectively. Furthermore, the RMSE was reduced by 16.2% and 21.7% for ASTGCN compared to HA and GRU, respectively. For diabetes drugs, ASTGCN achieves 22.9% lower MAE, and 20.3% lower RMSE, compared with HA, and achieves 31.5% lower MAE, and 28.6% lower RMSE, compared with GRU.

As prescription demand datasets are often collected on a monthly basis, they tend to be relatively small in size. In the above experiments, we noted that for the traditional neural network XGBoost, a small dataset inevitably brings about overfitting problems. However, for graph neural networks, due to its mechanism of parameter sharing and connection sparsity, it can better exploit the information of the data, which allows it to reduce the risk of overfitting.

TABLE I
HYPERLIPIDEMIA PRESCRIPTION DEMAND PREDICTION

| Model | MAE | RMSE |
|---|---|---|
| HA | 0.1216 | 0.1474 |
| ARIMA | 0.1280 | 0.1651 |
| XGBoost | 0.1839 | 0.2375 |
| GRU | 0.1194 | 0.1502 |
| STGCN | 0.0991 | 0.1299 |
| ASTGCN | 0.0955 | 0.1234 |

TABLE II
DIABETES PRESCRIPTION DEMAND PREDICTION

| Model | MAE | RMSE |
|---|---|---|
| HA | 0.1241 | 0.1557 |
| ARIMA | 0.1718 | 0.2233 |
| XGBoost | 0.1885 | 0.2370 |
| GRU | 0.1398 | 0.1738 |
| STGCN | 0.1139 | 0.1465 |
| ASTGCN | 0.0957 | 0.1241 |

*4) Performance at Different Locations:* Different regions exhibit distinct spatial-temporal correlation patterns, and each model has its own advantages in different patterns. In this section, we analyze the performance of each model across various prediction locations, and the corresponding results are shown in Table III.

The results clearly indicate that the accuracy of graph neural network-based models, STGCN and ASTGCN, surpasses that of other models in over 52.9% of the regions. Among them, ASTGCN, which performed best, achieved the best performance in 118 locations. Even where the rest of the models perform best, ASTGCN showcased comparable accuracy, with minimal deviation.

TABLE III
LOCATIONS WHERE EACH MODEL PERFORMS THE BEST

| Model | # of locations | % of locations |
|---|---|---|
| HA | 37 | 10% |
| ARIMA | 47 | 12.4% |
| XGBoost | 44 | 11.6% |
| GRU | 50 | 13.4% |
| STGCN | 83 | 21.8% |
| ASTGCN | 118 | 31.1% |

*5) Performance with Longer Prediction Window:* Long-term prediction is also an essential part of prescription demand prediction applications. In the experiment, we explored the ability of graph neural network-based models for long-term prediction by progressively increasing the output size with a one-month step and employing GRU

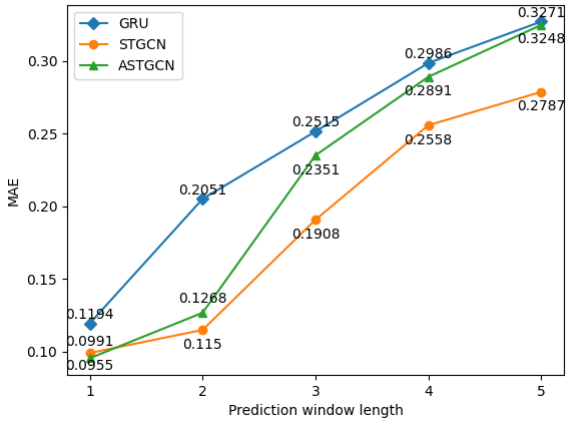


Fig. 1. MAE of each model under different prediction window length

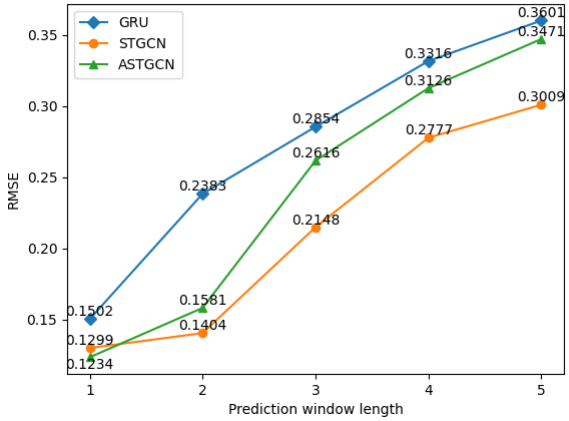


Fig. 2. RMSE of each model under different prediction window length

as a comparison. We report the results in Figure 1, 2. The results show that both MAE and RMSE increase with increasing prediction window length. This is due to the diminished temporal correlation of further future moments with the existing training data. However, both the ASTGCN and STGCN models, which leverage graph neural networks, consistently outperformed the GRU model. It is worth noting that while ASTGCN showcased better accuracy for short-term predictions, STGCN proved to be more effective for long-term forecasts.

## VI. Conclusion

In this paper we present the region-level prescription demand problem and investigate the spatial-temporal correlation of prescription data demand. The characteristics of low data volume, missing data, and outliers in prescription data put forward high requirements for the prediction model. To tackle these challenges, we analyze three types of models. Through a series of experiments, it is shown that graph neural network-based models outperform traditional forecasting models in terms of various evaluation metrics. This demonstrates their effectiveness in capturing and leveraging spatio-temporal correlations in the prescription data. As part of our future work, we intend to develop more accurate models that can effectively exploit correlations between data with different prescription requirements while better extracting spatio-temporal correlations from the data.

## References

[1] B. Wettermark, M. E. Persson, N. Wilking, M. Kalin, S. Korkmaz, P. Hjemdahl, B. Godman, M. Petzold, L. L. Gustafsson, and t. R. D. E. C. in Stockholm County Council, "Forecasting drug utilization and expenditure in a metropolitan health region," *BMC Health Services Research*, vol. 10, p. 128, May 2010.

[2] E. D. Kantor, C. D. Rehm, J. S. Haas, A. T. Chan, and E. L. Giovannucci, "Trends in Prescription Drug Use Among Adults in the United States From 1999-2012," *JAMA*, vol. 314, pp. 1818–1830, 11 2015.

[3] A. Bora, A. Nirali, C. Chaudhari, D. Gavade, and V. Vaishnav, "A time dependent epidemiology model for hospital resource management in usual scenarios and pandemic," in *ICDSMLA 2020* (A. Kumar, S. Senatore, and V. K. Gunjan, eds.), (Singapore), pp. 501–510, Springer Singapore, 2022.

[4] Y. Feng, J. Wang, Y. Wang, and S. Helal, "Completing missing prevalence rates for multiple chronic diseases by jointly leveraging both intra- and inter-disease population health data correlations," in *Proceedings of the Web Conference 2021*, WWW '21, (New York, NY, USA), p. 183–193, Association for Computing Machinery, 2021.

[5] S. Zhu, A. Bukharin, L. Xie, M. Santillana, S. Yang, and Y. Xie, "High-resolution spatio-temporal model for county-level covid-19 activity in the u.s.," *ACM Trans. Manage. Inf. Syst.*, vol. 12, sep 2021.

[6] 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," 2020.

[7] J. Gao, C. Xiao, L. M. Glass, and J. Sun, "Popnet: Real-time population-level disease prediction with data latency," in *Proceedings of the ACM Web Conference 2022*, WWW '22, (New York, NY, USA), p. 2552–2562, Association for Computing Machinery, 2022.

[8] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[10] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.

[11] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 922–929, 2019.

[12] J. Gao, R. Sharma, C. Qian, L. M. Glass, J. Spaeder, J. Romberg, J. Sun, and C. Xiao, "Stan: spatio-temporal attention network for pandemic prediction using real-world evidence," *Journal of the American Medical Informatics Association*, vol. 28, no. 4, pp. 733–743, 2021.

[13] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[14] S. Miller, A. El-Bahrawy, M. Dittus, M. Graham, and J. Wright, "Predicting drug demand with wikipedia views: Evidence from darknet markets.," in *Proceedings of the web conference 2020*, pp. 2669–2675, 2020.