Dynamics-Based Feature Augmentation of Graph Neural Networks for Variant Emergence Prediction

Majd Al Aawar [1], Srikar Mutnuri [2], Mansooreh Montazerin [1], Ajitesh Srivastava[1]

[1]University of Southern California
[2]University of Virginia
malaawar@usc.edu, nmp8rj@virginia.edu, mmontaze@usc.edu, ajiteshs@usc.edu

Abstract

During the COVID-19 pandemic, a major driver of new surges has been the emergence of new variants. When a new variant emerges in one or more countries, other nations monitor its spread in preparation for its potential arrival. The impact of the new variant and the timings of epidemic peaks in a country highly depend on when the variant arrives. The current methods for predicting the spread of new variants rely on statistical modeling, however, these methods work only when the new variant has already arrived in the region of interest and has a significant prevalence. Can we predict when a variant existing elsewhere will arrive in a given region? To address this question, we propose a variant-dynamics-informed Graph Neural Network (GNN) approach. First, we derive the dynamics of variant prevalence across pairs of regions (countries) that apply to a large class of epidemic models. The dynamics motivate the introduction of certain features in the GNN. We demonstrate that our proposed dynamics-informed GNN outperforms all the baselines, including the currently pervasive framework of Physics-Informed Neural Networks (PINNs). To advance research in this area, we introduce a benchmarking tool to assess a user-defined model's prediction performance across 87 countries and 36 variants.

Code — https://github.com/itssmutnuri/gnnvariants

Introduction

The COVID-19 pandemic presented an unprecedented global health crisis, severely affecting millions of people worldwide and demanding swift and effective responses from governments and healthcare providers (Organization 2023b). As the pandemic progressed, multiple variants of COVID-19 emerged, each possessing genetic mutations that can significantly impact transmissibility, virulence, and even vaccine efficacy. During much of the COVID-19 epidemic, the surge in cases and severe outcomes (hospitalizations and deaths) have been driven by the emergence of new variants (Wiemken et al. 2023). Consequently, monitoring the emergence and spread of these variants is crucial for devising appropriate public health measures and optimizing containment strategies (Organization 2023a). A critical factor driving the surge in a given region is the arrival time of the new variant (Markov et al. 2023). We can observe how a new variant, that has not appeared in region A yet, spreads in region B. We can study the spread in the region B to understand the properties of this new variant. If it is a highly transmissible or immune-evading variant (Lambrou et al. 2022), we can expect it to spread in the region A eventually. However, precisely when it would happen in region A remains unknown. Making a good prediction of arrival time would lead to more effective preparation and resource management.

We focus on the problem of predicting the arrival of a new variant in a given region provided that it has appeared somewhere else. Due to testing delays and the fact that not all cases are genomically analyzed, it is difficult to pin down when a variant arrives. Therefore, we consider measuring the delay to reach a certain proportion of prevalence.

Figure 1 shows the proportions of different variants over time in the United Kingdom and Sweden. We can see that there is often a clear delay (see 20.Alpha.V1 and 21.J.Delta) in arrival of the variants in Sweden. Furthermore, a variant may start to spread, but quickly get dominated by another variant before it reaches a significant proportion of the circulating cases (e.g., 21.A.Delta in Figure 1). Therefore, we reformulate the problem as the following:

Problem 1 (Emergence Delay Prediction) *Given the proportion of a variant in regions $A_1, A_2, \ldots$, predict **when** the variant will reach a proportion of θ in region B, provided that this variant has not yet reached region B.*

Here, the term "proportion" refers to the fraction of new cases created by the variant under consideration out of total new cases. Note that this is a regression problem – we seek to predict a quantity, which is the delay between the current date and the date of reaching the desired threshold. Further, we wish to perform this prediction for a given region before any sample of the variant of interest is observed there.

Prior works have focused on the problem of predicting the prevalence of a variant when it has already arrived in the region of interest (Sun et al. 2020; Hu et al. 2020). Specifically, logistic regression turns out to provide a good estimate of the prevalence over time (Shah et al. 2021; Palaniappan, V, and David 2022). However, these techniques require the variant to have a non-zero proportion already in the region of interest. To the best of our knowledge, no prior work ex-

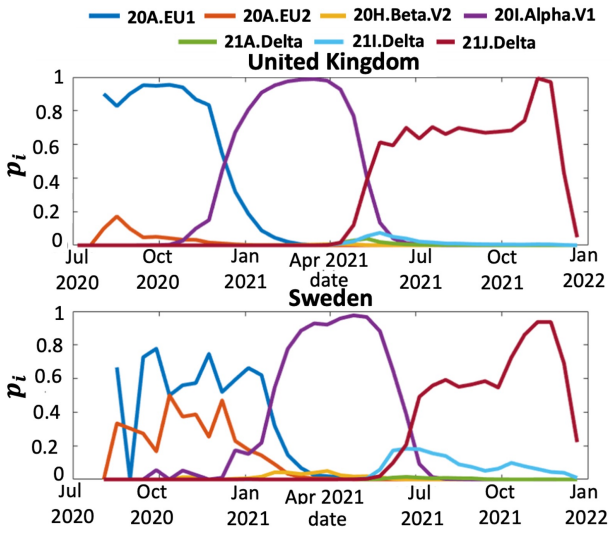


Figure 1: Plots showing the prevalence of a few different variants of COVID-19 over time in the UK and Sweden. The beginning of the spread of new variants can differ across countries by several weeks. This can be seen when examining "20I.Alpha.V1" and "21J.Delta" variants in the provided plots, showing their appearance in the UK several weeks before reaching Sweden.

ists predicting when the delay reaches a certain prevalence even when there is zero prevalence in the region.

When a new variant emerges with evolutionarily favorable properties, it can only be transferred to a different region through a host (in the case of COVID-19 – a human). This encourages the idea of using an underlying network of mobility to address the proposed problem. Since reaching a certain prevalence may depend on other currently circulating variants, their dynamics (how fast one variant can spread over others) also play a role. Therefore, we propose a variant dynamics-informed Graph Neural Network (GNN) that utilizes a network of mobility and features inspired by variant dynamics to solve the proposed problem. This is different than typical Physics-Informed Neural Networks (PINNs) where dynamics act as a regularization for the loss (Raissi, Perdikaris, and Karniadakis 2019a). Here, we show that our approach of constructing appropriate features results in lower errors compared to the PINNs approach.

Our key contributions represent a novel effort in addressing the challenge of predicting the emergence of COVID-19 variants at a global level, and in doing so establishing a new benchmark for evaluating Delay Prediction on the "CoVariants" dataset (Hodcroft 2021). We expect that this benchmark will help build the capacity to predict arrival times of the future variants of COVID-19 and other outbreaks. More specifically, our contributions are as follows:

1. We develop novel adaptations of GNNs that account for complex inter-dependencies between countries using GNNs while incorporating variant delay dynamics at the node level. To the best of our knowledge, we are the first to derive these delay dynamics for variants. Experiments demonstrate that our approach leads to superior results compared to several Machine Learning (ML) methods, including the currently pervasive framework of PINNs that integrate dynamics in the loss function.
2. We make our evaluation pipeline publicly available so that it can be used to evaluate any user-defined PyTorch model. This will advance public health research and enable public health experts and policymakers to leverage this benchmark pipeline for improvement of their models and consequently, for enhancing global health outcomes.

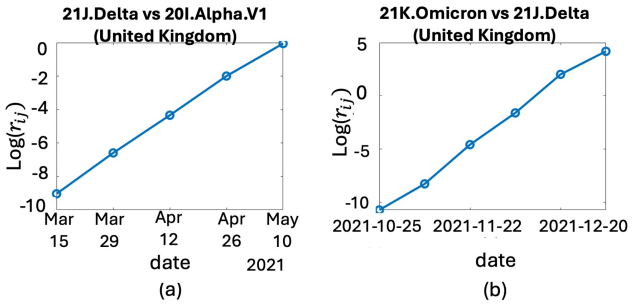


Figure 2: Sample semi-log plots of (a) 21J.Delta vs 20I.Alpha.V1 and (b) 21K.Omicron vs 21J.Delta for United Kingdom from data in the CoVariants dataset.

Related Work and Background

Variants and Variant Dynamics

In the context of infectious diseases, the term "variant" refers to a version of a virus with some changes in its genetic material, known as genome (for Disease Control and Prevention 2023). These changes happen through genetic mutations and can affect the virus's characteristics, like how easily it spreads, how severe the illness it causes is, and whether it can evade the immune system or not. Variants can be categorized in multiple ways based on genetic differences, and the significance of these differences can vary. Some categorizations focus on a common ancestor, while others may specifically highlight mutations in key regions of the virus's genome. In our study, we rely on the categorizations provided by the CoVariants dataset (Hodcroft 2021).

Furthermore, the proportions of different variants can be analyzed over time. Some studies have observed that these proportions follow a straight line when plotted on a semi-logarithmic scale. They leverage this fact to understand the dynamics of variant prevalence in populations (Beesley et al. 2023). We also observe this in our data as shown in Figure 2. However, to the best of our knowledge, existing studies do not derive any dynamics capable of predicting delays in the emergence of variants between multiple regions/countries.

Graph Neural Networks and Epidemics

Many efforts were made to develop forecasting models capable of predicting the progression and future trends of the COVID-19 pandemic (Cramer et al. 2022; Sherratt et al.

2023; J et al. 2020). However, these forecasting approaches have primarily focused on predicting the spread of the virus rather than predicting delays in variant emergence. Predicting the timing of variant-specific outbreaks can significantly enhance public health preparedness and response strategies, allowing for timely interventions such as targeted vaccination campaigns and tailored public health measures (Organization 2023a).

Many Deep Learning (DL) techniques have been explored for this purpose, including GNNs which help capture the graph effects of the spread of infectious diseases over time across regions (Scarselli et al. 2008; Davahli et al. 2021; Gao et al. 2021; Panagopoulos, Nikolentzos, and Vazirgiannis 2021; Ganesan and Subramani 2021). Other methods like PINNs (Seo, Meng, and Liu 2019; Seo and Liu 2019; Raissi, Perdikaris, and Karniadakis 2019b), agent-based models (Chopra et al. 2022, 2021), and a combination of spatio-temporal networks with location-aware features (Deng et al. 2019), including mobility data (Kapoor et al. 2020) have also been used.

Instead of only predicting virus spread, as others have done, we aim to predict when a new variant will emerge in a region given that it has emerged elsewhere. This involves deriving dynamics of variant spread and integrating these insights into a GNN. Notably, our strategy revolves around crafting relevant features rather than altering the loss function, providing a distinctive and effective approach.

Methodology

Dynamics-based Model

First, we develop an understanding of the dynamics that play a role in the spread of a new variant and how we can utilize them to make informed predictions. We start by extracting each variant's *global growth rate* from observed data. This value provides an indication of the dominance of the variants over other competing ones. Subsequently, we construct linear models based on the dynamics of infectious diseases and the derived growth rates to predict delays between regions, establishing this *Dynamics-based Model* as our baseline for *Delay Prediction*. Finally, we consider different approaches in which we can incorporate some disease dynamics using GNNs that are chosen for their ability to intuitively model spatial dynamics, with the goal of enhancing the prediction accuracy of our model.

Throughout the text, we use the terms "prevalence" and "prevalence ratio", where the "prevalence", p, is the proportion of cases of a given variant, and the "prevalence ratio" is defined as $r_{ij} = \frac{p_i}{p_j}$, where p_i is the prevalence of the variant i, and p_j is the prevalence of the most prevalent variant, j, that is not i. In our experiments, a variant is considered dominant if it reaches an r_{ij} value of at least $\frac{1}{3}$. We define this value as the *dominance threshold*, θ.

Variant Growth Rates We first derive the dynamics of two competing variants in one region. While logistic regression has been widely used to fit these dynamics, for completeness, we show in (Aawar et al. 2024) that the prevalence ratio between two variants i and j can be approximated by equation:

$$\ln(r_{ij}(t)) = \frac{t}{b}\ln(\beta_{ij}) + \ln(r_{ij}(0)) = S_{ij}t + C \quad (1)$$

Here, constant b is the mean serial interval – the average time between two successive infections, and β_{ij} is the ratio of transmission rates between the two competing variants. Parameter S_{ij} can be referred to as the relative growth rate of variant i over variant j (Srivastava 2023). Consider a scenario in two regions A and B — A has variant i emerging over the previously dominant variant j, but variant j never reached region B where variant k is dominant. This makes it difficult to assess the potential impact of variant i on region B when we only know S_{ij}. We address this by noting that the parameter S_{ij} should ideally be independent of the region. However, this may not hold due to the simplifying assumptions to derive Equation 1. Therefore, we introduce **region-specific relative growth rates** $S_{ij}^{(q)}$ for each region q and find the growth rate of any variant i relative to a fixed variant 0, specifically, the original COVID-19 variant. This is independent of any region and is defined as the global growth rate of variant i, denoted as S_i (Aawar et al. 2024). Finally, based on the global growth rates, we define **global relative growth rates** S_{ij} as $S_i - S_j$.

Now, for a pair of regions X and Y, the delay τ can be shown to be :

$$\tau = \frac{S_{kj}}{S_{ij}S_{ik}}\ln\left(\frac{\theta}{1-\theta}\right) - \frac{S_{kj}}{S_{ij}S_{ik}}C'_X + \frac{C'_X - C'_Y}{S_{ij}} \quad (2)$$

where θ is a given prevalence and C'_X and C'_Y are the parameters for linear regression. This expression serves as the foundation for our *Dynamics-based Model* baseline. The full derivation can be seen in (Aawar et al. 2024).

This is a pairwise linear model that estimates the delay between only two regions rather than predicting the delay for a region with respect to all other regions. Algorithm 1 details how we compute the date at which a variant arrives in a region Y. We calculate the median of the outputs from all pairwise models from regions X_q to region Y. We evaluate the performance of all the models (X_q, Y) until the current date and select the top three performing source regions among $X_1, X_2, \ldots$. Notably, the best performance here refers to the lowest error in predicting the delay. We identify three source regions that produced the lowest error in predicting the delays at $t-1$. Then, we take the median of the predictions produced by these selected models at time t which represents the delay at time t for the new variant to appear in region Y.

Dynamics-Informed GNN

Key Idea: We observe that global growth rates S_i play a crucial role in the dynamics. Furthermore, based on Equation 1, under some assumptions, the logarithm of the prevalence ratio r_{ij} grows linearly with time. We hypothesize that using $\ln(r_{ij})$ and S as features simplify the underlying patterns to be learned by an ML algorithm. An ML algorithm may also complement any violations of the assumptions made in the *dynamics-based model*. We propose a GNN-based network for this problem as GNNs are adept at capturing spatial

Algorithm 1: Dynamics-based Arrival Date Computation

for X_q in $\{X_1, X_2, \dots\}_t$ **do**
    Calculate and store delay predictions for (X_q, Y) until the current date, t.
if $t \neq 0$ **then**
    Select top 3 X_q with the best performance at $t-1$
else
    Select all source regions $\{X_1, X_2, \dots\}_t$
Find the median of delay predictions for selected models
return Median delay as the estimated arrival date for the variant in region Y at time t

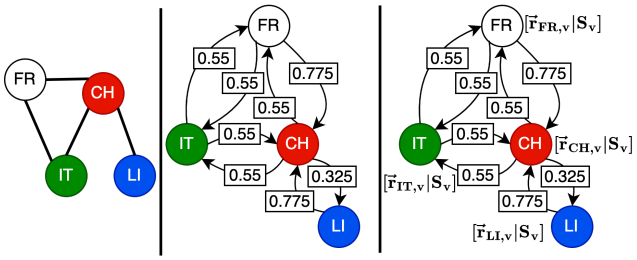


Figure 3: Graph creation process on sample subgraph $\mathcal{G}$: (i) First, we construct our nodes and edges based on country connectivity (ii) Next, we account for temporal variations in the relations between countries, i.e. the edges, to get $\mathcal{G}_t$ (iii) Finally, we find the variant specific features to get our final graph $\mathcal{G}_{t,v}$. Both (ii) and (iii) are for t = July 2nd, 2021.

relationships, making them well-suited for problems where geographical proximity between regions, such as countries, plays a crucial role. Additionally, the problem induces a natural graph structure, where countries can be represented as nodes with edges denoting some relationship between the countries. Furthermore, we explore different techniques to incorporate disease dynamics into the GNN in an effort to capture the complex interactions and patterns associated with the spread of infectious diseases.

Graph Construction The graph construction is illustrated in Figure 3. First, we represent each country as a node in the graph, and edges are established to represent relationships between the countries. There are 87 countries in which at least one variant appears and becomes dominant in the CoVariants dataset (Hodcroft 2021). The dataset provides the state of variants and mutations of interest for COVID-19 from August 2020 through October 2023. It has a bi-weekly resolution for 36 variants starting from 20A.EU1 to 23F.Omicron. While not shown in Figure 3, self-loops are included to account for internal transmissions.

Next, we consider the temporal aspect of the problem. The graph may evolve over time to capture changing relationships or influences among countries. This is introduced into our graph as dynamic edge weights based on border control data in the "OxCGRT" dataset (Hale et al. 2021). This dataset is from January 2020 to January 2023 measuring the variations in government responses using their COVID-19 Government Response Stringency Index, which is a simple additive score of nine indicators ranging from school closures to vaccination policies. Of these indicators, we picked international travel closure controls since they best capture the emergence and cross-border transmission of variants. These values are encoded in an ordinal scale as follows: 0 - no measures, 1 - screening, 2 - quarantine on high-risk regions, 3 - ban on high-risk regions, 4 - total border closure. The recorded values were then scaled down to a range between 0.1 and 1, with 1 indicating no border restrictions and 0.1 representing complete border closure. Note that 0.1 was chosen over 0, acknowledging the possibility of some mobility between connecting countries due to necessary trade or border crossings via open land routes. Given that our CoVariants data extends beyond January 2023, we extend OxCGRT by assuming that all border restrictions have remained unchanged since Jan 8, 2023.

To obtain the full graph, we create multiple copies of the above graph representing regions and inter-connectivity – one per variant at a given timestamp t. This approach allows us to leverage variant-specific dynamics as our node features. At t, each node c has features consisting of a time series of the logarithm of prevalence ratios r_{ij} of a present variant i and the most prevalent variant $j \neq i$, spanning T time steps. This results in a vector $\vec{r}_{c,i} = \left[\ln\left(r_{ij}^{t-T}\right), \ln\left(r_{ij}^{t-T+1}\right), \dots, \ln\left(r_{ij}^{t}\right)\right]$. The corresponding S_i value of the variant i is then concatenated to form each node's final feature vector $[\vec{r}_{c,i}|S_i]$.

Feature Augmented Graph Convolutional Network (FA-GCN) To extract both temporal and geographical information of the data, we propose a hybrid architecture summarized in Figure 4a. In this model, we employ a simple Gated Recurrent Unit (GRU) (Cho et al. 2014) to extract temporal information of the variant transmissions in T time steps. We temporarily remove the growth rate S_i (which is not time-dependent) from node features and embed the 1-dimensional (1D) time-dependent features (prevalence ratios) into a 2D latent feature vector. Note that in this configuration, we embed the features separately for each country, treating the nodes as distinct input samples grouped into a single batch, resulting in a 3D latent feature vector. The 3D latent feature vector is then flattened to obtain a final 2D embedding, which serves as input for a one-layer Graph Convolutional Network (GCN). The Leaky ReLU activation function introduces non-linearity after the GCN layer. Additionally, two GraphNorm(Cai et al. 2021) layers are applied after the GRU and GCN layers for normalization, and dropout is incorporated after the GCN layer for regularization. Finally, the embedded features are fed into a Fully Connected (FC) layer to predict delay until emergence for each country.

Physics Informed GCN (PI-GCN) For comparison, we use a PINN-inspired (Raissi, Perdikaris, and Karniadakis 2019a) approach to make predictions based on dynamics for the Delay Prediction task. This model utilizes a 2-layer GCN architecture, incorporating dropout and two GraphNorm layers which are applied to the output of each GCN layer, along with Leaky ReLU activation functions. Typically this GCN would be trained using a Mean Squared Error (MSE) loss, but we modify this loss function to steer our GCN towards

dynamic predictions made by the dynamics-based model.

The adjusted loss function is given by:

$$\text{MSE}_{\text{adjusted}} = \frac{1-p}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \frac{p}{N}\sum_{i=1}^{N}(\hat{y}_i - \dot{y}_i)^2 \tag{3}$$

where N is the number of countries, $\hat{y}_i$ is our model's output, y_i is the ground truth, and p is a hyperparameter ranging from 0 to 1, signifying the influence of $\dot{y}_i$, the output from the dynamics-based model. This modification ensures that, during training, the model is penalized more when its prediction deviates from the linear model's prediction, providing a form of dynamics-informed regularization. The best p value was found to be 0.1 by grid search.

Training Procedure

All training and validation of the models were performed retrospectively. This means that the model is trained and validated at each time step using only the "observed" data available up to that time step, denoted as d. For each variant, the retrospective algorithm iterates through all the biweekly data associated with that variant, training the model only on the weeks that have already passed (i.e., the observed data). Note that the pre-processing is also done retrospectively, accounting for any smoothing, calculation of the S values, or interpolation in the case of the dynamics-based model. This ensures that the approach is applicable in a prospective setting, where no data from the future is available.

The train/validation split is also done temporally to ensure that validation is performed only on the most recently observed data. This aims to achieve a better fit to the latest data, enhancing the model's predictive performance. An Early Stopper monitors the validation loss and halts the training when overfitting is detected, signified by a constant and substantial increase in the validation loss. The Early Stopper saves the best-performing weights, which are reused in the next iteration. This practice eliminates the need for the model to start training from scratch, leveraging knowledge gained from previous iterations.

We use MSE, or its adjusted version, as the loss function. Another crucial aspect is how to handle data related to variants that either do not appear in a given country or appear but fail to become dominant, i.e. rapidly diminish. In such cases, where regression targets would be undefined (potentially infinite), we address this challenge by creating a mask. This mask is employed to conceal nodes corresponding to these specific variants/country pairs, ensuring they do not contribute to our training loss. Essentially, our graph models are not trained on these nodes.

Experiments

Since we are the first to attempt solving this problem, we provide a benchmarking tool for the community, detailed in (Aawar et al. 2024). It presents the comprehensive pipeline employed for retrospective training, validation, and testing of diverse models. Its design prioritizes user-friendliness, allowing seamless integration of new PyTorch models or direct utilization of scikit-learn models through a configuration file. Users only need to specify whether the model requires graph data, leaving the pipeline to handle the rest. Additionally, the data pre-processing is encapsulated in a separate function, providing users with the flexibility to make modifications. The evaluation is also done retrospectively. This means that at time t, the training is performed using data only available until then to predict the delay for a given variant in a given country where it has not been seen yet. Considering the biweekly nature of the data, the target is a multiple of 2 weeks. Furthermore, evaluations exclude countries where a variant has already become dominant or never achieved dominance.

Pre-processing

The initial pre-processing stage aligned the country names across diverse datasets, with an emphasis on inclusion based on data availability. Only countries present in all datasets were retained, ensuring a standardized set for subsequent analyses. For each country, the prevalence of each variant was calculated at each timestep. Variants with a prevalence of less than 5% were disregarded in our analysis. To calculate the prevalence ratios, the two variants with the highest prevalence were identified for each timestep. For each variant on that particular day, the prevalence ratio was calculated. Finally, variants that did not reach a prevalence ratio of 0.2 or those that were present in less than three countries were filtered out. These steps ensured that only variants with a significant presence and wide geographic distribution were included in the subsequent analysis. All subsequent pre-processing steps were conducted retrospectively. This implies that as time progresses, more data is used for computing the global growth rates (S) and for training.

Results

In this section, we present our key results for Delay Prediction. We previously noted that prior research has concentrated on predicting the prevalence of a variant after its arrival in a specific region, with logistic regression demonstrating effectiveness in estimating its prevalence over time. However, as logistic regression can manage this task relatively easily, our evaluation focused solely on forecasting a variant's prevalence before its arrival in a specific region. In other words, during evaluation, we excluded cases where a variant is present in a region but has not yet reached θ.

Numerous models were tested and configured as benchmarks to test against the performance of our proposed method. Other implementations include a trivial *Mean Model*, the baseline *Dynamics-based Model*, a Decision Tree, a Multilayer Perceptron (MLP), a GRU, and a GCN without an adjusted loss function. In brief, the Mean Model predicts delays based on the average delays of all prior variants between all the countries in which the variant appeared and our intended country. The Decision Tree, MLP, and GRU regression models make use of T extra features which consist of the average $\vec{r}_{c,i}$ of neighboring countries transforming the dataset into a matrix of size $N \times (2T+1)$. Here, $N = D \times V \times C$ is the total number of samples, D is the total number of retrospective dates, V is the total number

Model	MedMedAE	MedMAE
Mean Model	3.2	3.87
Dynamics-based Model	1.86	2.9
Decision Tree	3.48	3.43
MLP	2.38	2.48
GRU	1.5	1.37
GCN	1.32	1.43
FA-GCN	**1.27**	**1.29**
PI-GCN	2.24	2.01

Table 1: Prediction Errors

of variants, and C is the total number of countries. The full details on these models can be found in (Aawar et al. 2024).

To measure the performance of the models, we use the defined Mean/Median Absolute Errors (MAE/MedAE) in (Aawar et al. 2024) and measure the performance over time by finding the mean and median of these errors as:

$$\text{MedMAE}_v = \text{median}(\text{MAE}_{v,t}),$$
$$\text{MedMedAE}_v = \text{median}(\text{MedAE}_{v,t}),$$

where t ranges from 1 to T_D, T_D being the total number of timesteps for which variant v circulates before reaching total dominance, i.e. it globally reaches all its targets and begins to disappear. Finally, the mean across all the variants v gets our two error metrics: MedMAE, and MedMedAE.

Table 1 presents the models' performance metrics for Delay Prediction. The results indicate that our proposed FA-GCN model outperforms all other models in terms of MedMedAE and MedMAE, showcasing its effectiveness in capturing temporal dependencies for Delay Prediction. The dynamics-based model also has a significantly worse MedMAE than MedMedAE. This suggests that the dynamics-based model is susceptible to outlier predictions for some country pairs, which skews its results. Additionally, we observe that the PI-GCN does not surpass its GCN counterpart, highlighting the unnecessary use of dynamics as a regularization of the loss when employing the correct features.

Discussion

Results Analysis

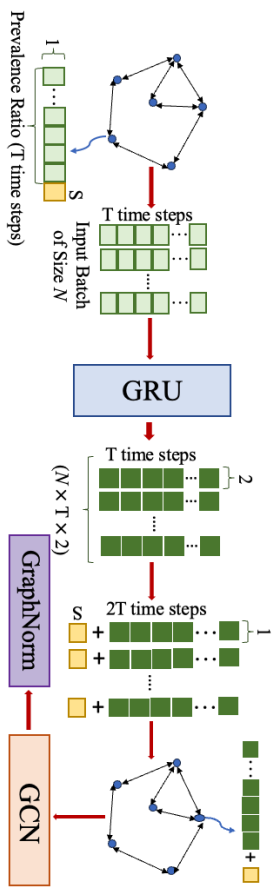
Figure 4b provides a detailed representation of the results for the baseline dynamics-based model and the FA-GCN model. The heatmaps illustrate the number of countries eligible for prediction at each timestamp. The numbers indicate the errors over time for each variant throughout their existence. Notably, several variants in the dynamics-based model are denoted by $*$, signifying instances of model or evaluation failure. These failures stem from three primary factors. Firstly, the linear model requires a minimum of two common variants between two countries to build an effective model. Secondly, the linear model can encounter significant challenges when variant j is identical in growth rate to variant k (i.e., $S_j = S_k$). In such cases, the linear model fits a plane solely along a fixed axis and will thus always predict some constant $\frac{C}{S_{ij}}$, as seen in Equation 2. But as time passes and variant j is no longer identical to variant k for another variant, this model fails drastically since it would still fit on a fixed axis. These scenarios were more prevalent in the early stages of the pandemic when only a few variants were circulating, leading to their exclusion from the dynamics-based model's training process. To ensure fair comparisons, all final errors across all models were calculated excluding these variants. Finally, there are cases where a variant appears in a country but fails to reach the threshold θ. These instances were disregarded during the evaluation of all models as they were considered trivial. The observations from Figure 4b give us two insights. Firstly, the primary source of error is associated with variants that persist for an extended duration, such as 21J.Delta or 23E.Omicron. Examining the color gradients, these variants appear to linger in only a few countries before spreading to others. In reality, variants typically do not endure for such prolonged periods without either being superseded by another variant or transmitting to additional countries (Chen et al. 2023; Health 2023). This suggests a potential issue in the data capturing these lingering variants. Secondly, despite this being a major source of error for FA-GCN, it exhibits overall robustness to outlier predictions, unlike the dynamics-based model as seen in Figure 4b. This, in turn, severely affects the results of the PI-GCN model.

Model	$[r\|S]$	r	$[p\|S]$	p	S
Decision Tree	3.48	3.0	**2.81**	3.0	2.86
MLP	**2.38**	2.67	2.69	2.55	2.76
GRU	**1.50**	1.50	1.95	2.32	n/a
GCN	**1.32**	1.50	1.78	2.32	2.38
FA-GCN	**1.27**	1.55	1.73	1.73	n/a

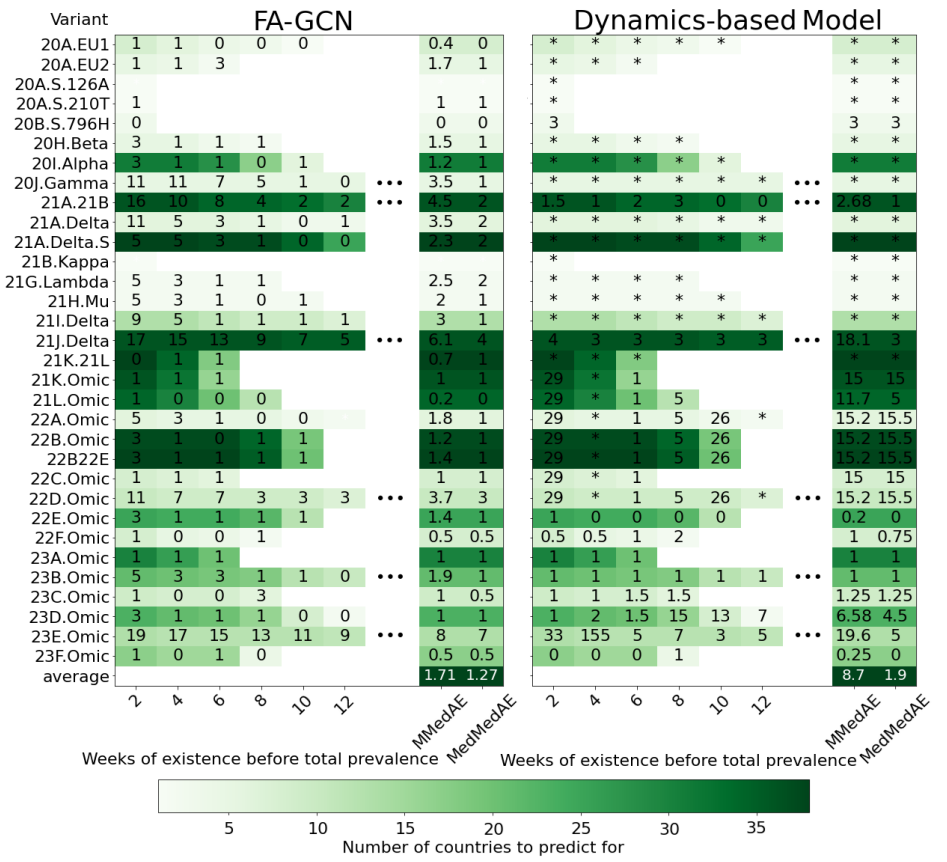
Table 2: Feature Ablations MedMedAE

Ablation Study: Dynamics

We examine the usefulness of integrating the extracted dynamics as inputs for our models. In this analysis, we train the same models using different sets of features. Specifically, we investigate the effectiveness of using S as a feature and the impact of using a time series of r. Where r excludes S as a feature, relying solely on the time series of prevalence ratios rather than the prevalence p. Additionally, we investigate the effectiveness of S by comparing $[p|S]$ to p and utilizing S as the only feature. From the results in Table 2, we observe that the $[r|S]$ combination outperformed all other configurations except for the case of the Decision Tree. There is also an improvement when utilizing r over p for the temporal-based models, with the GCN performing similarly and the others exhibiting poorer performance. This suggests that the temporal models effectively encode the latent information contained in the time series of r, while the simpler models struggle to do so. Additionally, inclusion of S as a feature enhances results across all cases. Combining S and r significantly improves performance in most cases, indicating a synergistic effect likely stemming from the linear relationship between the two, as previously discussed. It is important to note that the GRU and FA-GCN were not trained using just S since excluding the time series encoding



(a) FA-GCN Architecture



(b) The heat maps display both the errors and the number of countries being predicted for each variant at each timestep until they are no longer dominant anywhere

Figure 4: Visualization of FA-GCN architecture and heat maps for prediction errors.

reduces these models to an MLP and GCN, respectively.

Implementation

The benchmarking pipeline (Aawar et al. 2024) was implemented in Python. This was carried out on a machine equipped with a 32-core Intel(R) Xeon(R) Gold 5218 CPU running at 2.3 GHz and 64 GB of RAM. Focusing on the example of variant 23F.Omicron at a single timestep, which contains the majority of the data, and the heaviest model (FA-GCN), the average runtimes for different components were approximately 32 seconds for data pre-processing, 20 seconds for training, and 0.01 seconds for inference. The entire retrospective pipeline takes approximately 2 hours to complete. Notably, the most time-consuming aspect of the pipeline is the data pre-processing step, which could potentially be optimized further through memorization.

Limitations

One limitation of our approach in addressing the problem of Delay Prediction lies in handling situations where a variant might not even become dominant in a given country. To tackle this issue, we define the following problem:

Problem 2 (Dominance Prediction) *Given the prevalence (proportion) of a variant in regions $A_1, A_2, \ldots$, predict if the variant will ever reach a proportion θ in region B.*

In future work, we plan to address this issue by adopting a two-step approach. First, we solve Dominance Prediction, which aims to predict whether a variant will become dominant in a given region. If the variant is predicted to become dominant, we can then proceed to solve Delay Prediction.

Conclusion

We addressed the challenges of predicting variant delay across countries. The derivation of variant dynamics provided a theoretical foundation, which was used to engineer relevant features as well as a novel baseline dynamics-based model. We demonstrated that our dynamics feature-augmented GNN approach outperformed all other methods. Through comprehensive experiments and analysis, we demonstrated the effectiveness of our design choices, providing valuable tools for understanding and predicting the intricate relationships and connectivity patterns between nations. Furthermore, as we are the first to address the proposed problem, we provided a comprehensive benchmark and made our full pipeline available in an effort to facilitate research in the field.

Acknowledgements

This work was supported by the Centers for Disease Control and Prevention and the National Science Foundation under the awards no. 2135784, 2223933, and 2333494.

References

Aawar, M. A.; Mutnuri, S.; Montazerin, M.; and Srivastava, A. 2024. Global Prediction of COVID-19 Variant Emergence Using Dynamics-Informed Graph Neural Networks. *arXiv preprint arXiv:2401.03390*.

Beesley, L. J.; Moran, K. R.; Wagh, K.; Castro, L. A.; Theiler, J.; Yoon, H.; Fischer, W.; Hengartner, N. W.; Korber, B.; and Del Valle, S. Y. 2023. SARS-CoV-2 variant transition dynamics are associated with vaccination rates, number of co-circulating variants, and convalescent immunity. *eBioMedicine*, 91: 104534.

Cai, T.; Luo, S.; Xu, K.; He, D.; Liu, T.-Y.; and Wang, L. 2021. GraphNorm: A Principled Approach to Accelerating Graph Neural Network Training. arXiv:2009.03294.

Chen, J.; Gu, C.; Ruan, Z.; and Tang, M. 2023. Competition of SARS-CoV-2 variants on the pandemic transmission dynamics. *Chaos, Solitons; Fractals*, 169: 113193.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chopra, A.; Gel, E.; Subramanian, J.; Krishnamurthy, B.; Romero-Brufau, S.; Pasupathy, K. S.; Kingsley, T. C.; and Raskar, R. 2021. DeepABM: scalable, efficient and differentiable agent-based simulations via graph neural networks. *arXiv preprint arXiv:2110.04421*.

Chopra, A.; Rodríguez, A.; Subramanian, J.; Quera-Bofarull, A.; Krishnamurthy, B.; Prakash, B. A.; and Raskar, R. 2022. Differentiable agent-based epidemiology. *arXiv preprint arXiv:2207.09714*.

Cramer, E. Y.; Huang, Y.; Wang, Y.; Ray, E. L.; Cornell, M.; Bracher, J.; Brennen, A.; Castro Rivadeneira, A. J.; Gerding, A.; House, K.; Jayawardena, D.; Kanji, A. H.; Khandelwal, A.; Le, K.; Niemi, J.; Stark, A.; Shah, A.; Wattanachit, N.; Zorn, M. W.; Reich, N. G.; and Consortium, U. C.-. F. H. 2022. The United States COVID-19 Forecast Hub dataset. *Scientific Data*.

Davahli, M. R.; Fiok, K.; Karwowski, W.; Aljuaid, A. M.; and Taiar, R. 2021. Predicting the Dynamics of the COVID-19 Pandemic in the United States Using Graph Theory-Based Neural Networks. *International Journal of Environmental Research and Public Health*, 18(7): 3834.

Deng, S.; Wang, S.; Rangwala, H.; Wang, L.; and Ning, Y. 2019. Graph message passing with cross-location attentions for long-term ILI prediction. *arXiv preprint arXiv:1912.10202*.

for Disease Control, C.; and Prevention. 2023. CDC Variant Classifications. https://stacks.cdc.gov/view/cdc/105817. Accessed: 2025-02-16.

Ganesan, S.; and Subramani, D. 2021. Spatio-temporal predictive modeling framework for infectious disease spread. *Scientific Reports*, 11(1): 6741.

Gao, J.; Sharma, R.; Qian, C.; Glass, L. M.; Spaeder, J.; Romberg, J.; Sun, J.; and Xiao, C. 2021. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *Journal of the American Medical Informatics Association*, 28(4): 733–743.

Hale, T.; Angrist, N.; Goldszmidt, R.; Kira, B.; Petherick, A.; Phillips, T.; Webster, S.; Cameron-Blake, E.; Hallas, L.; Majumdar, S.; et al. 2021. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature human behaviour*, 5(4): 529–538.

Health. 2023. A Timeline of COVID-19 Variants. https://www.verywellhealth.com/covid-variants-timeline-6741198. Accessed: 2025-02-16.

Hodcroft, E. B. 2021. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. https://covariants.org/. Accessed: 2025-02-16.

Hu, H.; Du, H.; Li, J.; Wang, Y.; Wu, X.; Wang, C.; Zhang, Y.; Zhang, G.; Zhao, Y.; Kang, W.; et al. 2020. Early prediction and identification for severe patients during the pandemic of COVID-19: a severe COVID-19 risk model constructed by multivariate logistic regression analysis. *Journal of Global Health*, 10(2).

J, B.; D, W.; Deuschel, J.; Görgen, K.; Ketterer, J.; Gneiting, T.; and Schienle, M. 2020. The German and Polish COVID-19 Forecast Hub. https://kitmetricslab.github.io/forecasthub/. Accessed: 2025-02-16.

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

Lambrou, A. S.; Shirk, P.; Steele, M. K.; Paul, P.; Paden, C. R.; Cadwell, B.; Reese, H. E.; Aoki, Y.; Hassell, N.; Zheng, X.-Y.; et al. 2022. Genomic surveillance for SARS-CoV-2 variants: predominance of the Delta (B. 1.617. 2) and Omicron (B. 1.1. 529) variants—United States, June 2021–January 2022. *Morbidity and Mortality Weekly Report*, 71(6): 206.

Markov, P. V.; Ghafari, M.; Beer, M.; Lythgoe, K.; Simmonds, P.; Stilianakis, N. I.; and Katzourakis, A. 2023. The evolution of SARS-CoV-2. *Nature Reviews Microbiology*, 21(6): 361–379.

Organization, W. H. 2023a. Tracking SARS-CoV-2 variants. https://www.who.int/activities/tracking-SARS-CoV-2-variants. Accessed: 2025-01-30.

Organization, W. H. 2023b. World Health Organization COVID-19 Dashboard. https://covid19.who.int/. Accessed: 2025-01-30.

Palaniappan, S.; V, R.; and David, B. 2022. Prediction of epidemic disease dynamics on the infection risk using machine learning algorithms. *SN computer science*, 3(1): 47.

Panagopoulos, G.; Nikolentzos, G.; and Vazirgiannis, M. 2021. Transfer Graph Neural Networks for Pandemic Forecasting. arXiv:2009.08388.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019a. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378: 686–707.

Raissi, M.; Perdikaris, P.; and Karniadakis, G. E. 2019b. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378: 686–707.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1): 61–80.

Seo, S.; and Liu, Y. 2019. Differentiable physics-informed graph networks. *arXiv preprint arXiv:1902.02950*.

Seo, S.; Meng, C.; and Liu, Y. 2019. Physics-aware difference graph networks for sparsely-observed dynamics. In *International Conference on Learning Representations*.

Shah, D. A.; De Wolf, E. D.; Paul, P. A.; and Madden, L. V. 2021. Accuracy in the prediction of disease epidemics when ensembling simple but highly correlated models. *PLOS Computational Biology*, 17(3): 1–23.

Sherratt, K.; Gruson, H.; Grah, R.; Johnson, H.; Niehus, R.; Prasse, B.; Sandmann, F.; Deuschel, J.; Wolffram, D.; Abbott, S.; and et al. 2023. Predictive performance of multi-model ensemble forecasts of covid-19 across European nations.

Srivastava, A. 2023. The variations of SIkJalpha model for COVID-19 forecasting and scenario projections. *Epidemics*, 45: 100729.

Sun, J.; Chen, X.; Zhang, Z.; Lai, S.; Zhao, B.; Liu, H.; Wang, S.; Huan, W.; Zhao, R.; Ng, M. T. A.; et al. 2020. Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. *Scientific reports*, 10(1): 21122.

Wiemken, T. L.; Khan, F.; Puzniak, L.; Yang, W.; Simmering, J.; Polgreen, P.; Nguyen, J. L.; Jodar, L.; and McLaughlin, J. M. 2023. Seasonal trends in COVID-19 cases, hospitalizations, and mortality in the United States and Europe. *Scientific Reports*, 13(1).