COVID-19 Prediction with Doubly Multi-task Gaussian Process

**Sooyon Kim[1], Yongtaek Lim[2], Sungjun Lim[1], Gyeongdeok Seo[1], Jihee Kim[1],
Hojun Park[3], Jaehun Jung[3*], Kyungwoo Song[1*]**

[1]Department of Statistics and Data Science, Yonsei University, Seodaemun-gu, Yonsei-ro 50, Seoul, 03722, Republic of Korea
[2]Bigglz Inc, 4, Eoeun-ro 51beon-gil, Yuseong-gu, Daejeon, 34139, Republic of Korea
[3]Artificial Intelligence and Big-Data Convergence Center, Gil Medical Center, Gachon University College of Medicine, 38-13, Dokjeomro 3beon-gil, Namdong-gu, Incheon, 21565, Republic of Korea
mulan98@yonsei.ac.kr, yongtaek.lim@bigglz.com, lsj9862@yonsei.ac.kr, sgd3565@yonsei.ac.kr, jihee_sta@yonsei.ac.kr,
bacojun127@gmail.com, eastside1st@gmail.com, kyungwoo.song@yonsei.ac.kr

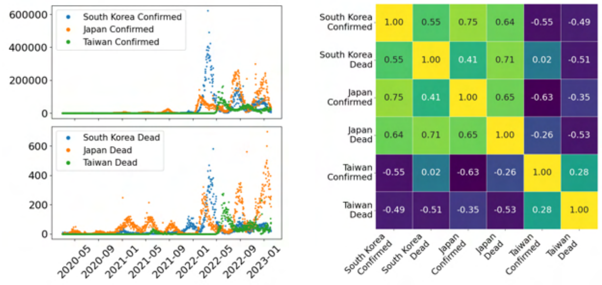
Abstract

This paper addresses a real-world multi-task prediction problem with time-series characteristics by proposing a novel Doubly Multi-Task Gaussian Process (DMTGP) model. Motivated by strong correlations between the number of confirmed cases and deaths, as well as between cases across the different countries, the model incorporates task-wise correlations to predict the number of COVID-19 patients, considering both task-specific (individual) and cross-task (shared) information to enhance overall performance. We constructed a database for three East Asian countries—Japan, South Korea, and Taiwan—and aim to simultaneously predict the number of confirmed cases and deaths in each country. To model the interactions among these countries, we employed a Transformer encoder layer to calculate cross-attention scores. Qualitative analysis of the attention score map demonstrates that our framework effectively captures the dynamic relationships between multiple nations over time. Our experimental results show that the DMTGP model outperforms other baseline models in handling doubly multiple tasks.

Introduction

Modeling the trend of infectious diseases is critical for understanding their spreading patterns and formulating effective public health responses. Among the numerous studies in the field, a particularly noteworthy effort focuses on the prediction of the number of confirmed COVID-19 cases and fatalities (dead cases). Given the highly infectious nature of COVID-19, accurate predictions of both confirmed and dead cases are essential for public health planning and intervention strategies.

In this study, we collected COVID-19 data from East Asian countries, specifically Japan, South Korea, and Taiwan, spanning three years. Our objective is to predict the number of confirmed cases and deaths simultaneously, as these tasks are highly correlated and can be inferred from the same data. The correlation coefficients between the number of confirmed cases and deaths in Japan, South Korea, and Taiwan are 0.72, 0.82, and 0.75, respectively. Additionally, strong correlations exist between the trends in different countries. Fig 1 shows the number of confirmed, dead cases from different tasks (nations) are highly correlated during specific periods. Fig 1(b) is the correlation plot of Fig 1(a) between July 1 and October 30 of 2022, which intuitively reveals similar infectious trends among neighboring, socially connected countries. These observations suggest that a joint modeling approach could be highly beneficial.



(a) The number of confirmed, dead patients of three countries

(b) Correlation Plot

Figure 1: (a) shows the number of confirmed and dead cases. (b) illustrates the correlation of confirmed, dead cases of three countries during the highlighted area of (a). Six targets are highly correlated according to the block-diagonal element of the matrix. Japan and South Korea show strong positive correlation, whereas Taiwan and the others show negative correlation.

Motivated by these strong correlations, we propose the Doubly Multi-task Gaussian Process (DMTGP) regression framework. To simultaneously predict two outcomes within the same country, we utilize multi-task Gaussian Process (MTGP) regression. This approach leverages the interdependence between confirmed cases and fatalities, enabling more accurate and robust predictions. Furthermore, we introduce a novel method that integrates data from multiple countries into a single predictive process. This is crucial since social issues, such as vaccination campaigns and quarantine measures, in one country can significantly influence the COVID-19 trends in neighboring countries. To model the interactions between the three countries, we employed an atten-

[*]Corresponding authors: Jaehun Jung and Kyungwoo Song.

tion module from the Transformer encoder layer. This allows our model to dynamically capture the complex dependencies and interactions between COVID-19 trends across the countries.

The main contributions of our work are summarized as follows:

- We propose an end-to-end Gaussian Process model, enhanced with neural network modules, to simultaneously predict confirmed cases and deaths in multiple countries.
- Our model incorporates the correlation between related countries into the prediction process by employing an attention mechanism, improving the accuracy and relevance of the predictions.
- We constructed the COVID-19 database system of East Asian countries with diverse features encompassing temporal, social, clinical, and climatological variables.

By combining MTGP and Transformer-based attention mechanisms, our approach aims to enhance the predictive performance and provide valuable insights into the progression of the pandemic in East Asia. This study not only contributes to the field of disease modeling but also offers a methodological framework that can be adapted to other regions and diseases.

Related Works

Disease Prediction with Machine Learning

The application of machine learning and deep learning methodologies to model disease trends and effects has seen significant advancements, particularly in the context of the COVID-19 pandemic. Various studies have demonstrated the efficacy of these techniques in predicting disease progression and outcomes (Shorten, Khoshgoftaar, and Furht 2021). The adjustment of compartmental Gaussian Process priors has been applied to effectively model the impact of lockdown policies on a global scale (Qian, Alaa, and van der Schaar 2020). This approach provided insights into how lockdown measures impacted the spread of COVID-19 across different regions. Recent work has introduced a robust, real-time prediction model designed to estimate the probability of in-hospital COVID-19 patients requiring mechanical ventilation (Zhang et al. 2022). Their end-to-end neural network model integrated MTGP to manage the irregular sampling rates in observational data. Additionally, the model employs a self-attention neural network to enhance the prediction task.

Multitask Learning

Multi-task learning (MTL) aims to enhance learning efficiency and prediction accuracy by jointly learning multiple objectives from a shared representation. This approach leverages commonalities across tasks to develop robust shared features. Balancing tasks during training is often achieved by assigning weights to the loss of each task, a strategy that has been explored in various previous works (Zhang and Yang 2021). The Gradient Normalization algorithm dynamically tunes gradient magnitudes to automatically balance training in deep multitask models (Chen et al. 2018). tasks during training. An end-to-end multi-task learning architecture enables task-specific feature-level attention within a single shared network (Liu, Johns, and Davison 2019). These modules facilitate the learning of task-specific features from the global features while simultaneously allowing feature sharing across different tasks.

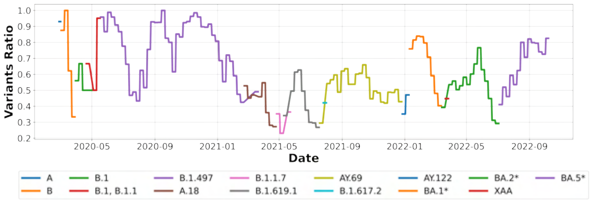


Figure 2: The most dominant COVID-19 variants changed over time. The plot includes data up to September 22, 2022.

Data Description

We collected the data ourselves, sourcing it from open-source and public datasets, covering the period from February 27, 2020, to February 8, 2023. The data encompasses three countries: South Korea, Japan, and Taiwan. It contains nine types of features, and we conducted a task to predict two of these—*Confirmation* and *Dead*—using the remaining seven features in a single country. In this section, we describe each feature and explain how we preprocessed the data. Table 1 presents the statistics of the database.

Temperature, precipitation, and humidity data were used as features derived from the weather information of each country. The predictions for data after February 8, 2023, were generated using data from two years prior. This approach allowed for a comprehensive analysis of how past weather conditions could impact future trends. Inoculation[1] refers to the number of vaccinated individuals. To standardize this metric across different countries, the vaccination rate was used by dividing the number of vaccinated individuals by the population of each country. Periods before the first reported vaccination were interpolated as having zero individuals. For dates beyond the observation period, interpolation was performed by fitting a linear function to the entire observed dataset, and any rates exceeding 100% were clipped at 100%.

COVID-19 variants data included information on the presence of diverse variants[2] and their respective proportions. In this study, since COVID-19 variant information was only available for South Korea among the three countries, the 28 variants identified in South Korea were used for all three countries. The dominant variants and their ratios, which varied by date, are presented in Fig 2.

The stringency index[3] is a metric primarily used to mea-

[1] The data of inoculation was collected based on https://github.com/owid/covid-19-data

[2] Global Initiative for Sharing All Influenza Data, GISAID provides the information about COVID-19 variants (Chen et al. 2022)

[3] This index derived from Oxford COVID-19 Government Response Tracker (OxCGRT) (Hale et al. 2021)

Table 1: Characteristics of self-collected data (N = 1,076). The data, encompassing South Korea, Japan, and Taiwan, includes nine features, and we predicted two of these—the proportion of confirmations and deaths among the population—using the remaining seven features within each country.

Nation	Japan				South Korea				Taiwan			
	Min	Mean	Std	Max	Min	Mean	Std	Max	Min	Mean	Std	Max
Temperature (°C)	-0.12	17.99	7.96	34.00	-15.13	13.99	10.75	33.00	3.00	23.95	5.82	35.0
Precipitation (mm)	0.00	6.82	23.33	280.65	0.00	3.78	11.69	106.48	0.00	6.72	15.50	230.41
Humidity (%)	23.00	62.12	15.27	99.75	15.00	60.40	15.53	99.12	39.00	70.96	9.73	98.00
Inoculation (%)	0.00	39.99	38.05	81.94	0.00	40.77	40.43	86.28	0.00	34.43	38.52	86.80
Stringency Index	17.67	33.41	5.61	44.09	11.11	35.75	16.84	74.16	3.70	22.08	12.76	65.85
Confirmation ($\% \times 10^2$)	0.00	2.49	4.30	23.73	0.00	5.54	12.75	11.98	0.00	4.03	8.48	69.24
Dead ($\% \times 10^4$)	0.00	0.52	0.73	5.54	0.00	0.62	1.23	11.20	0.00	0.69	1.53	11.60

sure the strictness of various measures implemented by the government during situations such as epidemics or emergencies. This index evaluates the types of actions implemented at specific times and assesses the intensity of those measures. The same interpolation strategy used for inoculation was employed.

Holiday data was utilized as a categorical value, assigning a value of 1 for public holidays and weekends and 0 for the rest of the days. Japan, South Korea, and Taiwan have 317, 336, and 346 holidays, respectively.

For the labels *Dead* and *Confirmation*, missing data were addressed as follows: Dates with initial missing values were imputed with zero. For gaps of 1 to 4 days, linear interpolation was utilized. For gaps of 5 to 7 days, the preceding days' values were averaged to fill the gaps. Additionally, label smoothing was applied to mitigate the impact of noise.

We generally employ min-max scaling for each feature and utilize linear interpolation to address null values within the dataset. We split the $N = 1076$ datasets in chronological order into $N_{\text{train}} = 860$ training sets and $N_{\text{test}} = 216$ test sets. The time feature was scaled within the training set as $\{i/N_{\text{train}}\}_{i=1}^{N_{\text{train}}}$. For the subsequent time periods in test sets, the scaling was applied as $\{1 + i/N_{\text{train}}\}_{i=1}^{N_{\text{test}}}$.

Methodology

Our proposed model is composed of a shared feature extractor network along with predictive MTGP heads, each corresponding to a different nation. We first define a *doubly multi-task setup*, where multiple response variables come from multiple different datasets. For obtaining shared features, we define multi-tasking as making predictions for different nations (datasets) simultaneously. Within a single nation, we also make predictions in a multi-task manner by simultaneously modeling the trends of confirmed and dead cases. To fully leverage nation-specific and cross-nation information, we adopt Deep Kernel Learning (DKL) and the attention mechanism in the Transformer encoder. These modules are trained together in an end-to-end fashion, as shown in Fig 3.

Our objective is to simultaneously forecast the occurrences of both confirmations and deaths in each country, taking into account time points and various features. The COVID-19 infection data exhibits not only temporal dependencies, due to its sequential nature and time-series attributes, but also spatial correlations because of the global pandemic context. Notably, events in one country can influence neighboring countries; for example, the impact of shutdowns, quarantines, or vaccination campaigns in South Korea can influence infection trends in Japan. To address these dependencies, we propose a novel architecture that leverages information from other countries when predicting COVID-19 cases for a specific country. Additionally, our proposed model incorporates convolution feature mapping and cross-attention mechanisms, enabling it to capture both temporal variations and spatial correlations.

Fig 3 illustrates the architecture of the proposed model. In the initial phase, we extract embeddings from the raw data of C countries using DKL as proposed by (Wilson et al. 2016). Subsequently, by employing a convolution kernel, both long-term and short-term temporal trends within features are captured, with the nature of these trends being contingent on the kernel size. Finally, to model inter-country correlations, we calculate cross-attention scores using the channel transformer. Our experiments demonstrate that the GP predictions improve significantly when using the rich feature representations, which are concatenated from shared and individual features.

Given C sets of correlated data from different countries, denoted as $X^{\text{raw}} = \{X_c^{\text{raw}}\}_{c=1}^{C}$, which are related and commonly influenced by major events, MTL learns set M of multiple tasks concurrently with a single shared network. Here each dataset has two tasks of predicting confirmed cases and dead cases.

Modeling Temporal Trend with Convolution Feature Map

When modeling infectious diseases, it is crucial to consider temporal pattern. According to Fig 1(a), the infectious trend appears in local manner. Motivated by previous studies that capture information over time at various scales, we aim to utilize multiple convolution layers to incorporate diverse temporal fluctuation trends (Chen and Shi 2021; Pan, Zhang, and Pu 2023).

After obtaining the non-linear embeddings $X^{\text{emb}} \in$

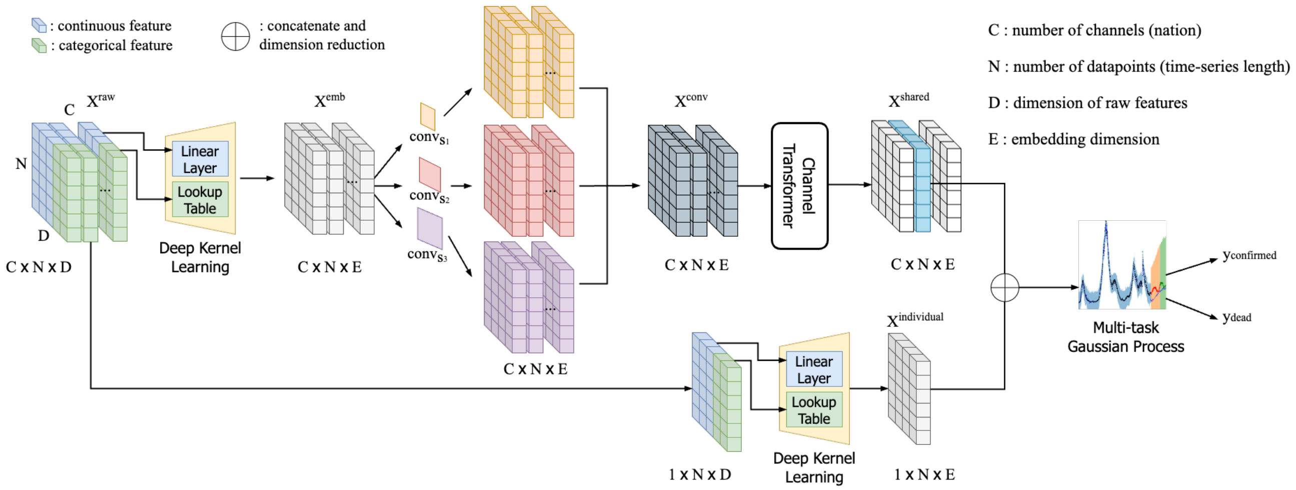


Figure 3: Illustration of the model architecture. The embedding X^{emb} is obtained raw data X^{raw} by DKL. Then, the model extracts local (short-term) trend via convolution map of different kernel sizes. Feature maps obtained from convolution are fused into X^{conv} by non-zero averaging. Averaged features go through Channel Transformer, yielding final shared feature X^{shared} with consideration to correlation between C countries and varying trend along time. Combined together with task-specific feature map $X^{\text{individual}}$, MTGP head is able to predict the infectious trend with cross-task, task-specific information.

$\mathbb{R}^{C\times N\times E}$ and $X^{\text{individual}} \in \mathbb{R}^{N\times E}$ through DKL, we elaborate X^{emb} through three convolution layers conv_{s_1}, conv_{s_2}, and conv_{s_3} to obtain $X^{\text{conv}_{s_1}}, X^{\text{conv}_{s_2}}, X^{\text{conv}_{s_3}}$. The kernel sizes of conv_{s_1}, conv_{s_2}, conv_{s_3} are s_1, s_2, s_3, respectively. In this study, we designate 7 for s_1, and 14 for s_2 to model the weekly, bi-weekly variability of infection trend. We employ zero padding to ensure consistent output dimensions when using convolution layers of varying sizes on time series data. Additionally, we employ causal 1-dimensional convolution following WaveNet (Oord et al. 2016), which enables the model to make predictions without having knowledge of future data points. After obtaining $X^{\text{conv}_{s_1}}$, $X^{\text{conv}_{s_2}}$, and $X^{\text{conv}_{s_3}}$, we compute their non-zero mean to ultimately obtain X^{conv}.

Spatial Correlation Modeling with Attention

The attention mechanism is widely employed in various forms to enhance feature representations. Originally proposed in the Transformer model for machine translation tasks, it excels in capturing intricate relationships between words at multiple levels (Vaswani et al. 2017). To achieve this in time-series data, the Transformer transforms the input elements X^{conv} into query (Q), key (K), and value (V) representations through a linear transformation, as defined in Eq 1.

$$Q = X^{\text{conv}}W_Q, K = X^{\text{conv}}W_K, V = X^{\text{conv}}W_V, \\ where\ W_Q, W_K, W_V \in \mathbb{R}^{E\times d} \quad (1)$$

Following Eq 2, the attention scores (A) are computed by taking the dot product between Q and K, scaled by $\sqrt{d}$. Here, d represents the dimension of the key, which in this case is identical to the original embedding dimension E. Finally, the attention-weighted sum of the values (V) is calculated to obtain the output feature representation $X^{\text{shared}} = \{X_c^{\text{shared}}\}_{c=1}^{C}$.

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), X^{shared} = AV \quad (2)$$

Using the attention mechanism, the Transformer can obtain a representation that reflects the attention between each nation in a single time point. In our task, we concatenate data of C countries, and then use the Transformer encoder layer to utilize the correlations between each country. In other words, we obtain embeddings, considering the spatial correlation between countries, using the attention matrix $A \in \mathbb{R}^{C\times C}$. We index the data of the c-th country X_c^{shared} from $X^{\text{shared}} = [X_{\text{train}}^{\text{shared}}; X_{\text{test}}^{\text{shared}}]$, which considers spatial correlations for all countries. Task specific embedding, which is obtained from data of single nation c and task-specific DKL module, is denoted as $X^{\text{individual}} = X_c^{\text{emb}}$. Finally, we concatenate $X^{\text{individual}} = [X_{\text{train}}^{\text{individual}}; X_{\text{test}}^{\text{individual}}]$ and X_c^{shared} and then reduce dimension to obtain final input $X_c = [X_{\text{train}}; X_{\text{test}}]$ for task-specific MTGP head. X_{train} and X_{test} have length N_{train} and N_{test}, respectively.

Deep Kernel Learning

Deep Kernel Learning transforms the inputs of GP kernel with a deep neural net (DNN) architecture (Wilson et al. 2016). Given predictors $\mathbf{x}, \mathbf{x}'$ and a base kernel $k^{\text{time}}(\mathbf{x}, \mathbf{x}')$, we transform the predictors as

$$k^{\text{time}}(\mathbf{x}, \mathbf{x}') \to k^{\text{time}}(g(\mathbf{x}, \mathbf{w}), g(\mathbf{x}', \mathbf{w})) \quad (3)$$

where $g(\mathbf{x}, \mathbf{w})$ is a non-linear mapping parameterized by $\mathbf{w}$, or embedding feature obtained by DNN architecture.

In our multi-task prediction dataset, embeddings of continuous, categorical variables are obtained from linear layers and lookup table, which are parametrized by $\mathbf{w}$. Obtained embeddings $g(\mathbf{x}, \mathbf{w})$ are denoted as X_c.

Time-series Forecasting with Multi-task Gaussian Process

Regarding the predictive process of MTGP, we define individual tasks to be the prediction of the number of confirmation and death for fixed nation c. GP conceptualizes time-series modeling as a regression problem, allowing for a flexible approach to modeling temporal dependencies and predictions.

Note that X_c is a dataset of nation c with cross-task, task-specific representation. Given separate train data and test data, and let the length of train data be N_{train} and the length of test data be N_{test}. The responses for $|M|$ tasks is denoted as $Y_c = [Y_{\text{train}}; Y_{\text{test}}] \in \mathbb{R}^{N \times |M|}$. We aim to predict unobserved response-values Y_{test} given Y_{train}.

For inputs of different timestamps $\mathbf{x}, \mathbf{x}' \in X_{\text{train}}$, we model the trend of tasks $m, m' \in M$ jointly with multi-task GP. The GP prior of task m at $\mathbf{x}$ is defined as follows:

$$\begin{gathered} f_m(\mathbf{x}) \sim \mathcal{GP}(\mu_m(\mathbf{x}), k(m, m', \mathbf{x}, \mathbf{x}')) \\ m, m' \in M, \ \mathbf{x}, \mathbf{x}' \in X_{\text{train}}, \end{gathered} \tag{4}$$

where μ denotes the mean function and $k(\cdot)$ denotes covariance function defined by kernel. In the nature of GP, covariance function and kernel is utilized interchangeably.

Multi-task Gaussian Process learns a shared covariance function based on input-dependent features and a flexible covariance matrix across tasks (Bonilla, Chai, and Williams 2007). To solve multi-task prediction problem, assume GP prior that additionally induces direct correlation between tasks as Eq 5. MTGP inherently introduces correlations among tasks by defining covariance function as follows:

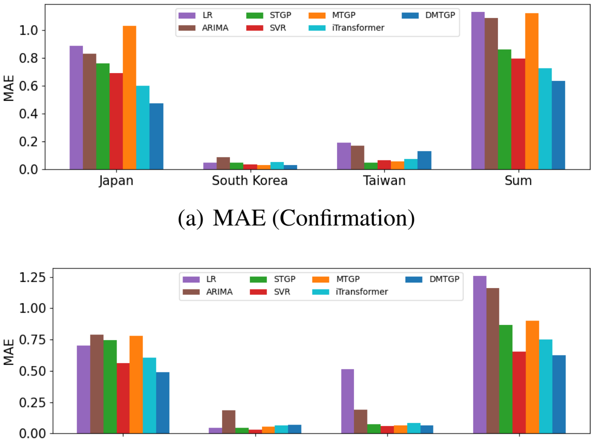
$$k(m, m', \mathbf{x}, \mathbf{x}') = k^{\text{task}}(m, m') \times k^{\text{time}}(\mathbf{x}, \mathbf{x}'), \tag{5}$$

Eq 5 denotes the covariance function, which models both task-wise correlation and correlation between inputs along the timestamp. $K^{\text{task}}_{M,M}$ and $K^{\text{time}}_{X_{\text{train}}, X_{\text{train}}}$ are Gram matrices of task kernel $k^{\text{task}}(m, m')$ and input kernel $k^{\text{time}}(\mathbf{x}, \mathbf{x}')$, respectively. Thus, entire covariance for $|M|$ tasks and train data X_{train} becomes $K^{\text{task}}_{M,M} \otimes K^{\text{time}}_{X_{\text{train}}, X_{\text{train}}} \in \mathbb{R}^{|M|N_{\text{train}} \times |M|N_{\text{train}}}$. The $\otimes$ denotes the Kronecker product, which is a mathematical operation that constructs a block matrix from two input matrices, and is used in multi-task Gaussian Processes to efficiently represent the joint covariance matrix for multiple tasks and inputs. After training, inference is done by using GP formula for the mean and variance of the predictive distribution with the covariance function given in Eq 5. For example, the mean prediction on a single data-point in the test set $\mathbf{x}_{\text{test}}$ for task m is given by

$$\begin{aligned} \bar{f}_m(\mathbf{x}_{\text{test}}) &= (\mathbf{k}^{\text{task}}_m \otimes \mathbf{k}^{\text{time}}_{\text{test}})^T \Sigma^{-1} \mathbf{y}_{\text{train}} \\ \Sigma &= K^{\text{task}}_{M,M} \otimes K^{\text{time}}_{X_{\text{train}}, X_{\text{train}}} + D \otimes I \\ D &= diag([\sigma_1^2, \ldots, \sigma_{|M|}^2]) \in \mathbb{R}^{|M| \times |M|}. \end{aligned}$$

In Eq 6, Σ is an $|M|N_{\text{train}} \times |M|N_{\text{train}}$ matrix and σ_m^2 denotes the noise variance for task m. D is a diagonal matrix of size $|M| \times |M|$, where the diagonal elements are $\sigma_1^2, \ldots, \sigma_{|M|}^2$. $\mathbf{k}^{\text{task}}_m \in \mathbb{R}^{|M| \times 1}$ column of task m in task kernel, whereas $\mathbf{k}^{\text{time}}_{\text{test}} = [k^{\text{time}}(\mathbf{x}_{\text{test}}, \mathbf{x}_1), \ldots, k^{\text{time}}(\mathbf{x}_{\text{test}}, \mathbf{x}_{N_{\text{train}}})]^T \in \mathbb{R}^{N_{\text{train}} \times 1}$ is the column of covariance functions between test data point and train data points. Thus, the Kronecker product of two vectors become vector of length $|M|N_{\text{train}}$, which is $(\mathbf{k}^{\text{task}}_m \otimes \mathbf{k}^{\text{time}}_{\text{test}})^T \in \mathbb{R}^{|M|N_{\text{train}} \times 1}$. To make inference regarding multiple tasks, matrix of response vectors are flattened as follows:

$$\begin{aligned} \mathbf{y}_{\text{train}} &= \text{vec}(Y_{\text{train}}) \\ &= (y_{11}, \ldots, y_{N_{\text{train}}1}, y_{12}, \ldots, y_{N_{\text{train}}2}, y_{1|M|}, \ldots, y_{N_{\text{train}}|M|})^T \end{aligned}$$



(a) MAE (Confirmation)

(b) MAE (Death)

Figure 4: Performance of the proposed model and the baselines on each task measured using MAE.

Optimization

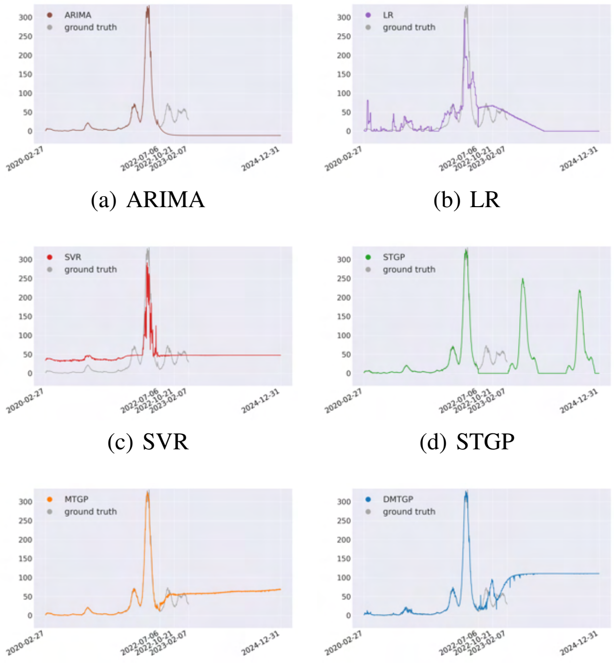
Learnable parameters are neural network parameters and kernel hyperparmeters which parametrize covariance function defined in Eq 5. Denote parameters in shared feature extractor modules as $\mathbf{w}$, and parameter of GP of nation c as θ_c. All the learnable parameters $\{\mathbf{w}, \theta_1, \ldots, \theta_C\}$ are optimized through maximizing marginal log likelihood (MLL) of multivariate Gaussian noise, which can be exactly computed by assuming Gaussian likelihood. Given C number of GPs of each country, parameters of GP kernels are optimized with respect to loss function $\mathcal{L}_c = \log p(Y_c | X_c, \theta_c, \mathbf{w})$, while parameters of shared modules across countries are optimized with respect to the sum of MLL of all country.

Experiments

We employ six different models for performance comparisons. Our problem is defined as doubly multi-task. In terms of the response variable, we predict the number of confirmed cases and dead cases for each country. Regarding the data, predictions are made for three countries within a single training and inference procedure. Thus, we aim to predict six tasks simultaneously: Japan confirmed cases, Japan dead cases, South Korea confirmed cases, South Korea dead cases, Taiwan confirmed cases, and Taiwan dead cases.

Table 2: The performance of the baseline models and the proposed model was evaluated with sums across three countries. The proposed multi-task prediction model demonstrated lower total Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) compared to the baseline models.

	MAE			RMSE		
MODEL	Conf	Dead	Sum	Conf	Dead	Sum
LR	1.1281	1.2573	2.3854	1.2688	1.4588	2.7526
ARIMA	1.0844	1.1618	2.2462	1.2287	1.3958	2.5886
STGP	0.8658	0.8653	1.7311	1.0417	1.0691	2.1108
SVR	0.7947	0.6542	1.4489	0.9376	0.8348	1.7724
MTGP	1.1180	0.8990	2.0170	1.2514	1.0991	2.3505
iTransformer	0.7261	0.7503	1.4763	0.8945	0.9430	1.8375
DMTGP	**0.6536**	**0.6225**	**1.2581**	**0.7219**	**0.7670**	**1.4889**



(a) ARIMA (b) LR

(c) SVR (d) STGP

(e) MTGP (f) DMTGP

Figure 5: Prediction trajectories of the baselines and the proposed method for the number of fatal(dead) patients for South Korea.

Baselines

ARIMA forecasts non-stationary time series data for a single task. Linear regression models relationships between variables for single-task forecasting. Support Vector Regressor forecasts two tasks per nation. GP predicts with uncertainties, where STGP forecasts either deaths or confirmed cases. MTGP predicts both deaths and confirmed cases simultaneously. The baselines ARIMA, LR, and STGP predict single tasks. SVR and MTGP serve as baselines predicting two tasks per country. iTransformer treats independent time series as variable tokens to capture multivariate correlations through attention mechanisms, and it employs layer normalization and feed-forward networks to learn representations of the series (Liu et al. 2023). iTransformer and DMTGP predicts all six tasks concurrently.

Quantitative Analysis

Table 2 compares the performance of various models on data from Japan, Korea, and Taiwan. To ensure balanced evaluation, we normalized the labels of both tasks (confirmed cases and deaths) to a 0-1 scale. Our model, DMTGP, consistently outperforms the others across all metrics. While SVR and iTransformer show competitive results, they fall short compared to DMTGP. SVR's lack of attention mechanisms limits its ability to capture correlations within the data. As illustrated in Fig 1(a), there are similar patterns in confirmations and deaths across the three nations, particularly a strong correlation between Japan and Korea, which our model captures through attention scores illustrated in Fig 1(b). iTransformer relies solely on label-based training, whereas DMTGP leverages multiple features, leading to more causal and interpretable predictions. Furthermore, Fig 4 compares the MAE of the number of confirmed cases and deaths across all countries. Our model demonstrates good performance with a significant margin in the Japan data compared to other models. Overall, DMTGP consistently demonstrates the best-aggregated performance across both metrics, making it the most effective model for this dataset despite the complexity of doubly multi-task scenario.

Qualitative Analysis

Modeling the Relationship Between Countries In this study, we investigate the influence of convolutional kernel size on the feature map and explore how our attention model captures relationships among countries. Specifically, we present the cross-attention score map derived from a feature map produced by a convolutional layer with a kernel size of 14 days (equivalent to two weeks). The results, depicted in Fig. 6, show matrices corresponding to selected days. The (i, j) element of the attention matrix represents the influence of the j-th nation (Key) on the i-th nation (Query), with a higher score indicating a stronger relationship. During the highlighted 14-day period, the confirmed cases and death cases exhibit similar patterns across all three countries, reflected in the relatively consistent attention scores influencing each other.

Sensitivity Analysis The choice of kernel is crucial for modeling trends with GP. To evaluate the impact of different kernels, we conducted experiments varying the combination of kernels across three countries. The kernels tested include the Spectral Mixture Kernel (SM), the Linear Kernel (L), and the Periodic Kernel (P). Kernels are multiplied and added to model complex patterns flexibly. For example, SMP+L kernel is the multiplication of SM and P kernel with addition to L kernel. In our experiments, we combined these kernels in various configurations to identify which combination yields the best predictive performance for each country. The combinations tested were SM+P, SM+L, SMP, and SMP+L. The performance was measured using MAE for confirmed cases, deaths, and total cases, as depicted in Fig

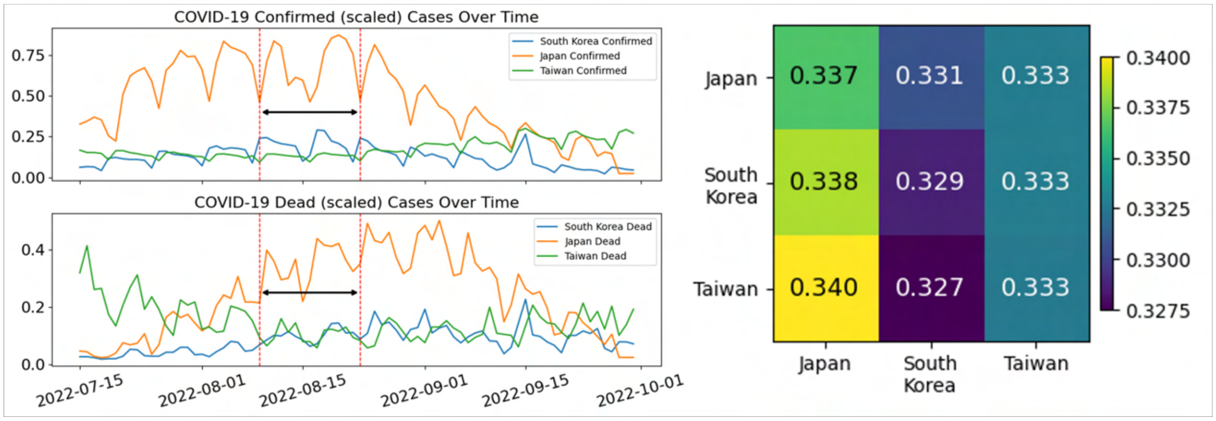


Figure 6: Attention score map (right) for August 23, 2022, derived from a feature map produced by a convolutional filter with a temporal window size of 14 days. The number of cases were scaled with minimal and maximal values.

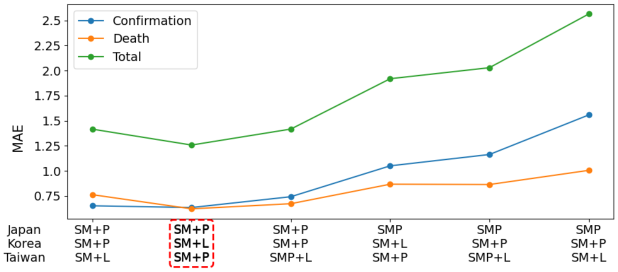


Figure 7: MAE according to the covariance kernels used to define the GP of each nation. Highlighted combination is selected DMTGP model.

7. By selecting the optimal combination of kernels for each country, we aim to improve the accuracy of our GP model in capturing the underlying trends in the data.

Conclusion

In this paper, we proposed a multivariate time-series forecasting framework, DMTGP, designed for a doubly multi-task setup where multiple prediction tasks exist across multiple task datasets. Specifically, we constructed and defined a doubly multi-task database for COVID-19, encompassing the number of confirmed and dead cases for three East Asian countries: Japan, South Korea, and Taiwan. To achieve accurate predictions, we employed MTGP, supported by a cross-task feature extractor neural network composed of DKL, Attention mechanisms, and various sizes of convolutional layers. Our model demonstrated superior predictive performance compared to baseline methods. The attention mechanism within our model effectively captured the correlations between different countries, enhancing the overall predictive accuracy. Furthermore, our approach seamlessly handled multiple task sequences within a single end-to-end training and prediction process. However, we acknowledge a limitation in our study: the sources of data vary among the countries, leading to some variables being naively imputed or interpolated. Despite this, we believe our model can be applied to a wide range of real-world time-series problems where multiple datasets are correlated.

/

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A4A3033874), and Yonsei University Research Fund of 2024-22-0123, and a grant (22183MFDS431) from ministry of food and drug safety in 2024.

References

Bonilla, E. V.; Chai, K.; and Williams, C. 2007. Multi-task Gaussian process prediction. *Advances in neural information processing systems*, 20.

Chen, C.; Nadeau, S.; Yared, M.; Voinov, P.; Xie, N.; Roemer, C.; and Stadler, T. 2022. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*, 38(6): 1735–1737.

Chen, W.; and Shi, K. 2021. Multi-scale attention convolutional neural network for time series classification. *Neural Networks*, 136: 126–140.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, 794–803. PMLR.

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.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1871–1880.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Pan, W.; Zhang, W.; and Pu, Y. 2023. Fractional-order multiscale attention feature pyramid network for time series classification. *Applied Intelligence*, 53(7): 8160–8179.

Qian, Z.; Alaa, A. M.; and van der Schaar, M. 2020. When and how to lift the lockdown? global covid-19 scenario analysis and policy assessment using compartmental gaussian processes. *Advances in Neural Information Processing Systems*, 33: 10729–10740.

Shorten, C.; Khoshgoftaar, T. M.; and Furht, B. 2021. Deep Learning applications for COVID-19. *Journal of big Data*, 8(1): 1–54.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wilson, A. G.; Hu, Z.; Salakhutdinov, R.; and Xing, E. P. 2016. Deep kernel learning. In *Artificial intelligence and statistics*, 370–378. PMLR.

Zhang, K.; Karanth, S.; Patel, B.; Murphy, R.; and Jiang, X. 2022. A multi-task Gaussian process self-attention neural network for real-time prediction of the need for mechanical ventilators in COVID-19 patients. *Journal of biomedical informatics*, 130: 104079.

Zhang, Y.; and Yang, Q. 2021. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12): 5586–5609.