Hi-COVIDNet: Deep Learning Approach to Predict Inbound COVID-19 Patients and Case Study in South Korea

Minseok Kim, Junhyeok Kang, Doyoung Kim, Hwanjun Song, Hyangsuk Min, Youngeun Nam, Dongmin Park, Jae-Gil Lee*
Graduate School of Knowledge Service Engineering, KAIST
{minseokkim,junhyeok.kang,doyo09,songhwanjun,hs_min,maradonam,dongminpark,jaegil}@kaist.ac.kr

ABSTRACT

The escalating crisis of COVID-19 has put people all over the world in danger. Owing to the high contagion rate of the virus, COVID-19 cases continue to increase globally. To further suppress the threat of the COVID-19 pandemic and minimize its damage, it is imperative that each country monitors inbound travelers. Moreover, given that resources for quarantine are often limited, they must be carefully allocated. In this paper, to aid in such allocation by predicting the number of inbound COVID-19 cases, we propose *Hi-COVIDNet*, which takes advantage of the geographic hierarchy. *Hi-COVIDNet* is based on a neural network with two-level components, namely, *country-level* and *continent-level* encoders, which understand the complex relationships among foreign countries and derive their respective contagion risk to the destination country. An in-depth case study in South Korea with real-world COVID-19 datasets confirmed the effectiveness and practicality of *Hi-COVIDNet*.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

COVID-19, coronavirus, pandemic, inflow infection prevention, deep learning, neural networks

ACM Reference Format:
Minseok Kim, Junhyeok Kang, Doyoung Kim, Hwanjun Song, Hyangsuk Min, Youngeun Nam, Dongmin Park, Jae-Gil Lee. 2020. Hi-COVIDNet: Deep Learning Approach to Predict Inbound COVID-19 Patients and Case Study in South Korea. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3394486.3412864

1 INTRODUCTION

World Health Organization (WHO) declared the COVID-19 outbreak a "pandemic" on March 11, 2020 [18]. The threat of the COVID-19 pandemic continues to increase worldwide through physical contact between people, thereby putting people in great danger. Owing to the high contagion rate of the disease, inbound infected patients might lead to a destructive pandemic, eventually paralyzing an entire country. To address this, most governments impose quarantines to monitor overseas inflow and prevent this infectious disease from entering across countries. Typically, epidemics come from abroad. The objective must be to allow a flow of uninfected travelers to avoid stopping economic activity.

South Korea's COVID-19 control and prevention measures, referred to as "K-Quarantine," received global praise. Regarding the control of overseas entrants, which is effective from April 1, 2020, all symptomatic entrants from abroad go through diagnostic tests at the airport. Those who test positive are transferred to a hospital or a community treatment center. The asymptomatic passengers (those who did not show symptoms of the disease) receive diagnostic tests at the airport if they come from Europe, and short-term foreigners are quarantined at government facilities. To operate this special entry procedure, the government must allocate resources, such as medical staff, diagnostic kits, and quarantine facilities, in advance, and then adjust the procedure to a potential new situation rapidly. Thus, it is very useful to precisely predict the number and trend of imported cases accurately.

In this paper, we focus on predicting the number of imported cases as precisely as possible for the near future (i.e., one or two weeks). The potential number of imported cases from a country can be represented as a function of inbound passengers arriving from that country and its respective degree of infection risk. Intuitively speaking, the number of imported cases from a country is proportional to the number of inbound passengers arriving from that country and the number of confirmed cases. However, the underlying relationship inside the function is too complex for simplification given that various factors change over time. For instance, a rapid spread of the disease within a country increases the pandemic risk of other countries immediately, while it also decreases inter-country interactions that are proportional to the infection risk. Simultaneously, a country tends to interact with nearby countries in the same continent more often than distant ones in other continents. Therefore, it is imperative for a prediction model to encompass such a complex, spatio-temporal relationship.

Following this intuition, we propose *Hi-COVIDNet*, a Hierarchical model that estimates the inbound COVID-19 cases from abroad based on the epidemic trend and inflection risk of foreign countries, effectively exploiting the power of deep neural Networks. To understand the innate nature of COVID-19, which spreads not only quickly but also unnoticeably, *Hi-COVIDNet* considers the temporal dependency of COVID-19 infections

*Jae-Gil Lee is the corresponding author.

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3412864

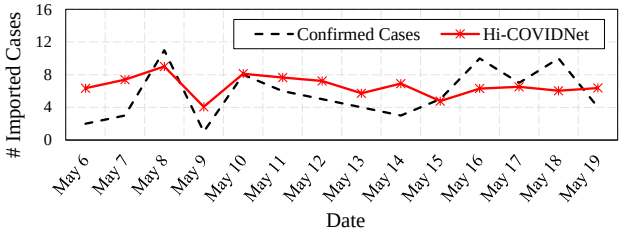


Figure 1: The predicted numbers for imported cases by *Hi-COVIDNet* together with the true numbers in South Korea.

country-wise through a recurrent neural network, followed by the incorporation of the risk factor and interaction information from each country. Furthermore, *Hi-COVIDNet* reflects the hierarchy of spatial contexts by aggregating foreign countries per continent. Our main contributions are summarized as follows:

- We propose *Hi-COVIDNet*, to the best of our knowledge, the first deep learning approach for estimating the upcoming number of imported COVID-19 cases.
- We exploit the geographic hierarchy as well as a hierarchical objective function to overcome a relatively short (i.e., approximately 1.5 months) period of data collection for COVID-19.
- We demonstrate the practicality and effectiveness of *Hi-COVIDNet* through a case study in South Korea. As an example, after training *Hi-COVIDNet* with datasets collected from March 22 through May 5, 2020, we predicted the number of imported cases for the period from May 6 through May 19, 2020 (two weeks). Figure 1 shows the prediction results together with the true numbers of daily imported cases. Note that the prediction results are very close to the true trend.

The rest of this paper is organized as follows. Section 2 summarizes the datasets used for evaluation. Section 3 proposes the architecture and methodology for *Hi-COVIDNet*. Section 4 presents the evaluation results. Section 5 reviews relevant studies for predicting the spread of epidemics. Finally, Section 6 concludes the paper.

2 DATA DESCRIPTION

In this section, we describe our large-scale data collection used to train *Hi-COVIDNet* for a case study in South Korea. The collection comprises *intra-country* and *inter-country* information. Table 1 summarizes the description of each dataset. All datasets, except for the roaming dataset, are publicly available on the Internet. They were collected from March 22 through May 5, 2020, to align the period with the information from Korea Centers for Disease Control and Prevention (KCDC).

2.1 Intra-Country Datasets

- **Confirmed Cases**: This dataset was provided by Johns Hopkins COVID-19 Resource Center[1]. It represents the number of daily confirmed cases and deaths per country. Additionally, we included their first and second derivatives to obtain the degree of COVID-19 infection speed per country.
- **Search Keywords**: This dataset was collected from Google Search Trend for the four keywords in Table 1, which represent the degree of the anxiety on the disease in each country.

[1]https://coronavirus.jhu.edu/map.html

Table 1: Summary of the datasets.

Dataset		Variable Description
Intra-Country Data	Confirmed Cases	(1) Date (2) Country (3) # of confirmed cases (4) First derivative of (3) (5) Second derivative of (3) (6) # of deaths (7) First derivative of (6) (8) Second derivative of (6)
	Search Keywords	(1) Date (2) Country (3)–(6) # of searches for "COVID-19," "COVID test," "Flu," and "Mask"
Inter-Country Data	International Roaming	(1) Date (2) Originating country (3) Total # of customers arriving in Korea
	Flights	(1) Date (2) Originating country (3) Total # of airlines arriving in Korea
	Imported Cases	(1) Date (2) Originating continent (3) Total # of imported cases in Korea

2.2 Inter-Country Datasets

- **International Roaming**: This dataset was provided by Korea Telecom (KT)[2], the second largest mobile carrier in South Korea. It contains its Korean customers returning to and from South Korea. We extracted the number of roaming entrants from each country per day to estimate the total daily inflow. Over 30% of Korean telecommunication users employ KT, which is sufficient to estimate the total number of Korean travelers.
- **Flights**: This dataset was collected from the airline information system[3]. It contains the number of daily cargo and passenger airlines arriving at Incheon Airport, the main international airport in South Korea. It was used as a rough estimate of the number of the entrants from abroad. Given that the roaming dataset covers only Korean travelers, this flight dataset was used together with the roaming dataset to cover all inbound travelers.
- **Imported Cases**: This dataset was collected by KCDC[4]. It contains the daily count of imported cases to South Korea, which are categorized by the originating continent. This daily count is used as the *label attribute* for training and testing. See Appendix A.1 for details.

3 METHODOLOGY: *HI-COVIDNET*

3.1 Overview

We propose a deep learning approach called *Hi-COVIDNet* that aims to predict the imported COVID-19 cases from abroad by learning the function of the degree of each country's infection risk and the amount of inbound passengers.

[2]https://corp.kt.com/eng/
[3]https://www.airport.kr/co/en/index.do
[4]http://ncov.mohw.go.kr/en

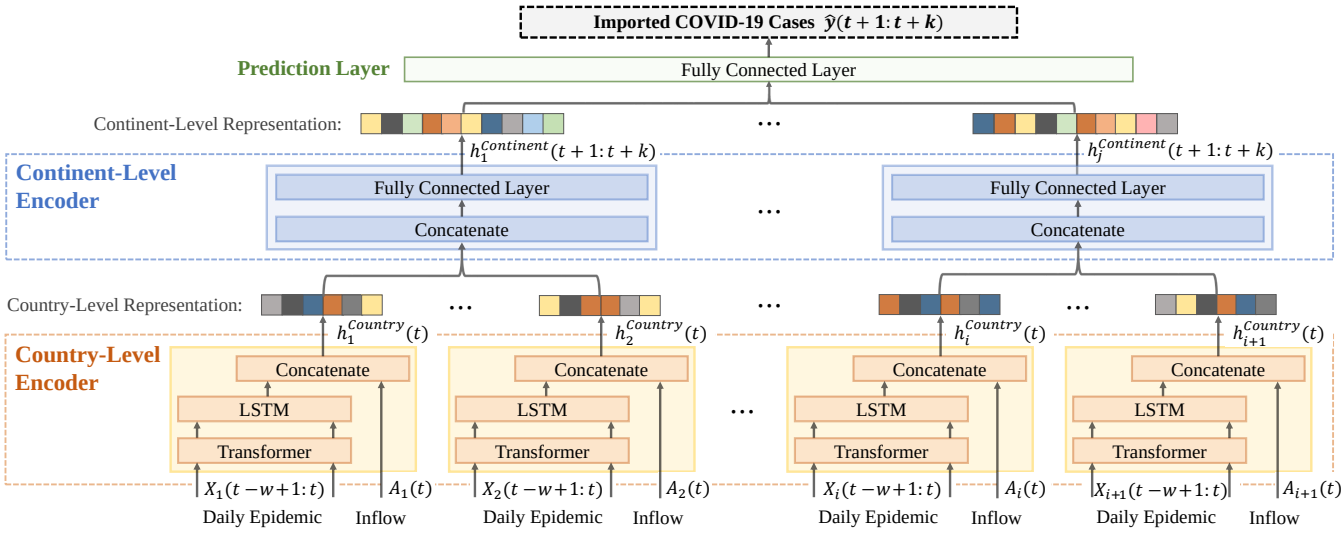


Figure 2: Two-level model architecture of *Hi-COVIDNet*.

Overall Architecture: Figure 2 depicts the *two-level* hierarchical architecture of our *Hi-COVIDNet* model, which mainly consists of the *country-level* encoder and the *continent-level* encoder:

1. **Country-Level Encoder**: This learns a hidden representation of both the infection risk in each country and the inflow trend from each country. The epidemic statistics (e.g., confirmed cases and deaths) and search keywords are provided for the former, and the roaming and flight statistics are provided for the latter. Then, all these inputs are concatenated to model the function of the two aspects.
2. **Continent-Level Encoder**: This aggregates the outputs of the country-level encoders belonging to the same continent, in that the spread of COVID-19 is greatly impacted by the neighboring countries. Thus, a hidden representation is generated per continent, which reflects those of the belonging countries.

At last, the prediction layer of *Hi-COVIDNet* returns the number of the imported COVID-19 cases to a destination country (i.e., South Korea) for upcoming k days, by combining the hidden representations of all continents and learning their respective contribution to the final output.

The notation required for describing Figure 2 in subsequent sections is summarized as follows[5]:

- $X_i(t)$ represents the epidemic statistics in the i-th country at day t. More specifically, $X_i(t)$ consists of the variables for *confirmed cases* and *search keywords* in Table 1, i.e., $X_i(t)$ = [# of confirmed cases, . . ., # of deaths, . . ., # of searches for "Mask"].
- $X_i(t-w+1:t)$ is the series of the epidemic statistics in past w days from day t, where w is the incubation period of COVID-19. More specifically, $X_i(t-w+1:t) = \langle X_i(t-w+1), \ldots, X_i(t)\rangle$.
- $A_i(t)$ represents the total inflow statistics from the i-th country to a destination country at day t. More specifically, $A_i(t)$ consists of the variables for *international roaming* and *flights* in Table 1, i.e., $A_i(t)$ = [# of inbound customers, # of inbound airlines].

[5]The variables for $X_i(t)$ and $A_i(t)$ can easily adapt to the datasets in hand.

- $y(t+1:t+k)$ is the series of the *true* numbers of daily imported cases to a destination country from day $t+1$ through day $t+k$, and $\hat{y}(t+1:t+k)$ is the series of the *predicted* numbers for the corresponding k days.
- $h_i^{Module}(\cdot)$ is the latent variable produced by *Module* for the i-th country or continent.

3.2 Country-Level Encoder

The country-level encoder consists of the Transformer [27] layer, the long short term memory (LSTM) [8] layer, and concatenate layer.

Transformer Layer: This layer highlights the significant period in $X_i(t-w+1:t)$ through the attention mechanism [27]. This layer is shown to be effective because a specific period can have a greater impact on the COVID-19 transmissions (see §4.3). For example, the 31st case in South Korea, discovered on February 18, 2020, participated in a worship service of a quasi-Christian cult called Shincheonji; since February 18, the number of confirmed cases literally exploded [21]. Overall, this layer is formulated by

$$h_i^{TM}(t-w+1:t) = Transformer(X_i(t-w+1:t); \Theta_i^{TM}), \quad (1)$$

where Θ_i^{TM} is the set of the parameters of the Transformer for the i-th country.

LSTM Layer: This layer intends to capture the temporal trend of the COVID epidemic, with the important period highlighted by the Transformer layer. The LSTM, a variant of recurrent neural networks (RNNs), is chosen here because it has achieved the state-of-the-art performance in various applications with sequence data [17]. To this end, we keep the output of the last (i.e., at day t) cell, which summarizes the entire sequence, in the LSTM. Overall, this layer is formulated by

$$h_i^{LSTM}(t) = LSTM(h_i^{TM}(t-w+1:t), \Theta_i^{LSTM}), \quad (2)$$

where Θ_i^{LSTM} is the set of the parameters of the LSTM for the i-th country. Accordingly, the output $h_i^{LSTM}(t)$ is the latent representation of the risk of infection in the i-th country at day t.

Concatenate Layer: This layer combines $h_i^{LSTM}(t)$, the output of the LSTM layer in Eq. (2), and $A_i(t)$. Recall that the former represents the risk of infection in the i-th country and the latter represents the total inflow from the i-th country to the destination country. Overall, this layer is formulated by

$$h_i^{Country}(t) = concat([h_i^{LSTM}(t), A_i(t)]). \quad (3)$$

Thus, the concatenated output $h_i^{Country}(t)$ is the hidden representation of the impact on the imported cases from the i-th country.

3.3 Continent-Level Encoder

The continent-level encoder, together with the hierarchical objective function (see §3.4), simply reflects the First Law of Geography, "everything is related to everything else, but near things are more related than distant things."

Encoder: This encoder consists of the concatenate layer and the fully-connected layer. The concatenate layer simply merges the outputs of the country encoders for the belonging countries in each continent. Then, a fully-connected layer is added to learn the relationship from the impact on the imported cases from the j-th continent at *day* t to the impact from the same continent during *upcoming* k *days*. Overall, this encoder is formulated by

$$h_j^{Continent}(t{+}1{:}t{+}k) = \phi(concat([\ldots, h_i^{Country}(t), \ldots]); \Theta_j^{\phi}), \quad (4)$$

where ϕ is a fully-connected network with one hidden layer[6] for the j-th continent and is parameterized by Θ_j^{ϕ} with the ReLU activation function. Therefore, the output $h_j^{Continent}(t{+}1{:}t{+}k)$ returns the expected impact of the imported cases for upcoming k days.

Prediction Layer: Finally, to estimate the *total* number, we further aggregate the outputs of the continent-level encoders by adding another fully-connected layer, formulated as

$$\hat{y}(t{+}1{:}t{+}k) = \psi([\ldots, h_j^{Continent}(t{+}1{:}t{+}k), \ldots]; \Theta^{\psi}), \quad (5)$$

where Θ^{ψ} is the set of the model parameters. The result $\hat{y}(t{+}1{:}t{+}k)$ is the predicted number of imported COVID-19 cases for the upcoming k days from day t.

3.4 Training Algorithm

Hierarchical Objective Function: To fully take advantage of the geographic hierarchy, our loss function considers (i) the prediction error on the number of the imported cases *from each continent* and (ii) the prediction error on the total number of imported cases. We follow the grouping of countries provided by KCDC and denote the set of continents by $C = \{$ China, Asia outside China, Europe, America, Africa, Oceania $\}$. The mean squared error (MSE) is used by default. As a result, the objective function is defined by

$$\begin{aligned}\mathcal{L} =& \beta \| (y_C(t{+}1{:}t{+}k) - \hat{y}_C(t{+}1{:}t{+}k)) \|_2^2 \\ &+ (1-\beta) \| (y(t{+}1{:}t{+}k) - \hat{y}(t{+}1{:}t{+}k)) \|_2^2,\end{aligned} \quad (6)$$

where $y_C(t{+}1{:}t{+}k)$, $\hat{y}_C(t{+}1{:}t{+}k) \in \mathbb{R}^{|C|\times k}$ denote the true and predicted numbers of continent-wise imported cases for upcoming k days. $\hat{y}_C = h_C^{Continent}$ when the dimensionality of the hidden variable is reduced to one. β is the hyperparameter for adjusting the weight of the continent-level error.

[6] Although this structure is well-known to approximate almost any continuous function [24], any other network structure such as a convolution neural network (CNN) [13] can be used instead.

Algorithm 1 *Hi-COVIDNet* Training

INPUT: $X_i(t)$, $A_i(t)$, and $y(t)$ (see §3.1); k days to predict
OUTPUT: Set of the optimal model parameters Θ_*

1: $\Theta^{TM}, \Theta^{LSTM}, \Theta^{\phi}, \Theta^{\psi} \leftarrow$ Initialize model parameters;
2: **for** *epoch* = 1 **to** *epochs* **do**
3: **for each** $t \in \{$ training days $\}$ /* Mini-batch */
4: /* COUNTRY-LEVEL ENCODER */
5: **for each** $i \in \{$ countries $\}$ **do**
6: Compute $h_i^{TM}(t)$ by Eq. (1);
7: Compute $h_i^{LSTM}(t)$ by Eq. (2);
8: Compute $h_i^{Country}(t)$ by Eq. (3);
9: /* CONTINENT-LEVEL ENCODER */
10: **for each** $j \in \{$ continents $\}$ **do**
11: Compute $h_j^{Continent}(t{+}1{:}t{+}k)$ by Eq. (4);
12: /* PREDICTION LAYER */
13: Estimate $\hat{y}(t{+}1{:}t{+}k)$ using Eq. (5);
14: /* MODEL UPDATE */
15: Compute the loss $\mathcal{L}$ by Eq. (6);
16: $\Theta^* \leftarrow \Theta^* - \alpha \nabla \mathcal{L}$;
17: **return** Θ_*;

Pseudocode for Reproducibility: Since *Hi-COVIDNet* is an end-to-end structure, we can optimize the entire model at once by deriving the partial derivatives during the minimization of the objective function in Eq. (6). The overall training procedure of *Hi-COVIDNet* is outlined in Algorithm 1. The pseudocode follows the presentation order in the previous sections. The forward propagation is performed for the country-level encoder (Lines 4–8), the continent-level encoder (Lines 9–11), and the prediction layer (Lines 12–13). Then, the hierarchical loss is calculated using the prediction result, and the all network parameters are updated by backpropagation (Lines 14–16). The training procedure stops when a given number of epochs elapses.

4 EVALUATION

To validate the superiority of *Hi-COVIDNet*, we conducted a case study to predict imported COVID-19 cases using *real-world* COVID-19 datasets in South Korea. Our evaluation was designed to support the followings: (i) *Hi-COVIDNet* provides *more accurate* prediction than the baselines (see §4.2); (ii) the main components of *Hi-COVIDNet* are indeed effective, and both data collections are helpful for prediction (see §4.3).

4.1 Experimental Setting

4.1.1 Datasets and Metric. The details of the datasets are presented in Section 2. We divided the entire period into the training set for March 22–May 5, 2020 and the test set for May 6–May 19, 2020. The window size w was set to be 14 according to the generally-known COVID-19 incubation period [30]. Note that our datasets were collected for a relatively short period at the time of this writing, compared with other datasets [5].

Table 2: RMSE comparison with the baseline methods.

Method	May 6–12 (k=7)	May 6–19 (k=14)
ARIMA	0.4931	0.6243
$LSTM_{sv}$	0.4600	0.4274
$LSTM_{mv}$	0.5188	0.4621
Hi-COVIDNet	**0.4373**	**0.4045**

We measured the prediction error using the root mean square error (RMSE) defined as

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(y_i(t{+}1{:}t{+}k) - \hat{y}_i(t{+}1{:}t{+}k)\right)^2}, \quad (7)$$

where $y(t{+}1{:}t{+}k)$ and $\hat{y}(t{+}1{:}t{+}k)$ are the true and predicted numbers, respectively, of imported COVID-19 cases for the upcoming k days from day t.

4.1.2 Algorithms and Implementation. Because the trend of imported cases is a kind of time-series, we compared *Hi-COVIDNet* with *two* popular time-series prediction algorithms, namely, the *ARIMA* [4] model and the *LSTM* [8] model. *ARIMA* only uses the variable of the number of imported cases for training and prediction. For the LSTM model, we used two variants: $LSTM_{sv}$ uses the single variable just like *ARIMA*, and $LSTM_{mv}$ uses the multiple variables, which are the same as those used by *Hi-COVIDNet*.

ARIMA was implemented with statsmodel[7], while the two *LSTM* variants and *Hi-COVIDNet* were implemented with PyTorch 1.2.0[8]. We ran all deep learning algorithms using an NVIDIA Tesla V100 GPU. In support of reliable evaluation, we repeated every test *five* times and reported the average. For reproducibility, we provide the source code and the data collection (except for the roaming dataset) at https://github.com/kaist-dmlab/Hi-COVIDNet.

4.1.3 Network & Training Configuration. We trained *Hi-COVIDNet* using the Adam [10] optimizer with a constant learning rate of 0.03 and a batch size of 1. The only hyperparameter β was set to be 0.5, which was the best value found in a grid $\beta \in [0, 1]$. For the compared algorithms, all hyperparameters were favorably set to the best values obtained by a thorough grid search. In all experiments, any regularization method (e.g., dropout [25], batch normalization [9], and weight decay [11]) was not applied.

4.2 Overall Comparison

We predicted the number of the imported cases for the upcoming 7 and 14 days. Table 2 shows the RMSE of the four algorithms for two prediction tasks. In addition, Figure 3 shows the prediction result in terms of the number of imported cases. Note that a prediction result is returned as a *real number*. Overall, *Hi-COVIDNet* achieved the lowest RMSE regardless of the duration of prediction, as shown in Table 2; and the predicted trend by *Hi-COVIDNet* followed the true trend most closely, as shown in Figure 3. These results indeed demonstrate the superiority and practicality of the hierarchical architecture of *Hi-COVIDNet*.

The RMSE of *ARIMA* was comparable to that of $LSTM_{sv}$ or $LSTM_{mv}$ when predicting upcoming 7 days, but *ARIMA* started

[7]https://www.statsmodels.org/stable/index.html
[8]https://pytorch.org/

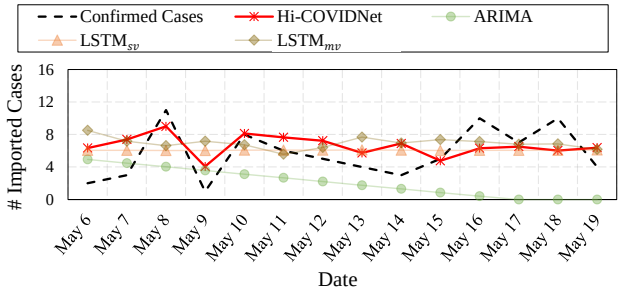


Figure 3: The predicted numbers for imported cases by *Hi-COVIDNet*, ARIMA, and two LSTM variants together with the true numbers in South Korea (best viewed in color). This figure repeats Figure 1 with the additional results of the three baselines.

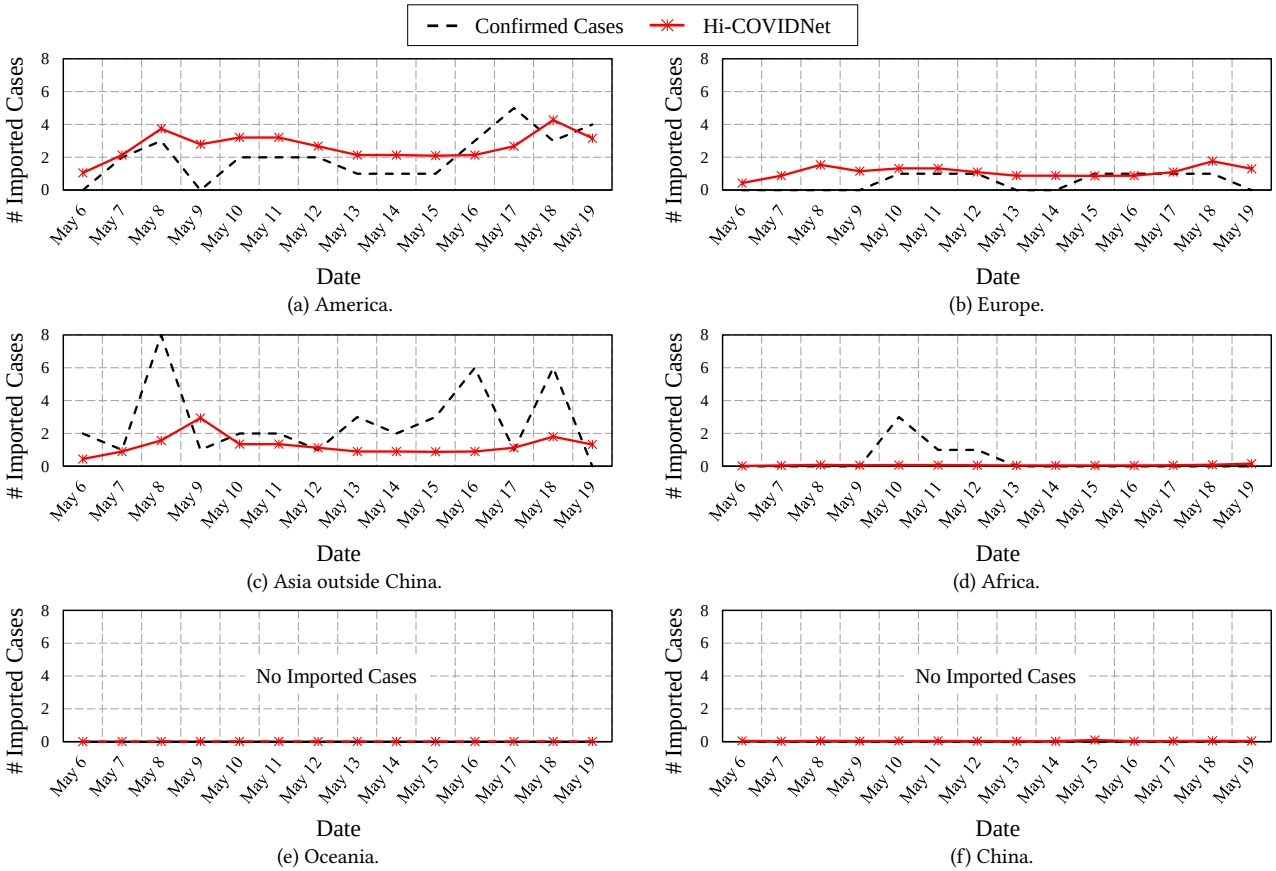
to be governed by its own bias when predicting upcoming 14 days, as shown in Figure 3. Thus, *ARIMA* was, in general, shown to be the worst. Meanwhile, the RMSE of $LSTM_{mv}$ was higher than that of $LSTM_{sv}$ despite more information, probably because naively adding information may even harm the model performance [6]. In addition, although exactly the same input was fed to *Hi-COVIDNet* and $LSTM_{mv}$, the RMSE of *Hi-COVIDNet* was much lower than that of $LSTM_{mv}$, thereby confirming again the advantage of our architectural design.

Continent-Wise Result: By virtue of the two-level architecture, *Hi-COVIDNet* is able to predict the number of imported cases *per continent*, as shown in Figure 4. The prediction results for each continent are interpreted as follows:

- America, Europe, and Asia outside China (Figure 4(a), Figure 4(b), and Figure 4(c)): *Hi-COVIDNet* continuously predicted a number of imported COVID-19 cases from these continents, because of both a high number of confirmed cases and a high number of inbound passengers. The ten countries with the highest number of confirmed cases (i.e., U.S., Brazil, Russia, U.K., Spain, Italy, France, Germany, India, and Turkey) are all from these continents as of May 30, 2020.
- Africa (Figure 4(d)): *Hi-COVIDNet* expected no imported COVID-19 cases because of a very low number of inbound passengers from Africa, though there were a few imported cases.
- Oceania (Figure 4(e)): *Hi-COVIDNet* expected no imported COVID-19 cases because new confirmed cases were very few in Oceania. In fact, there were no imported cases.
- China (Figure 4(f)): *Hi-COVIDNet* expected no imported COVID-19 cases because of very few inbound passengers from China. The Chinese government put travel restrictions, and the Korean government banned the entry from Wuhan. In fact, there were no imported cases.

4.3 Ablation Study

The outstanding performance of *Hi-COVIDNet* comes from various factors, including the hierarchical structure, the Transformer, and rich input data. To examine their respective effect, we excluded each of a few main components and additionally evaluated the accuracy of these variants. Table 3 shows the RMSE result of the *Hi-COVIDNet* variants.



(a) America. (b) Europe. (c) Asia outside China. (d) Africa. (e) Oceania. (f) China.

Figure 4: Predicted and true numbers for continent-wise imported cases in South Korea.

Table 3: Ablation study for the datasets and model components of *Hi-COVIDNet* (k=14).

Hi-COVIDNet Variant	RMSE
Hi-COVIDNet (w/o inter-country data)	0.6086
Hi-COVIDNet (w/o continent-level encoder)	0.5800
Hi-COVIDNet (w/o Transformer)	0.4543
Hi-COVIDNet	**0.4045**

Effect of Inter-Country Data: "*Hi-COVIDNet* (w/o inter-country data)" was not provided with the inter-country data (i.e., the roaming and flight datasets in Table 1). That is, $A_i(t)$ and the concatenate layer in Figure 2 were removed. Then, the RMSE was increased by up to 50.4%, which empirically verified that the inter-country data was effective to elicit the inflow trend.

Effect of Network Components: "*Hi-COVIDNet* (w/o continent-level encoder)" did not include the continent-level encoder in Figure 2. The RMSE was increased by up to 43.3%, and the two-level architecture did have a remarkable impact on the accuracy. On the other hand, "*Hi-COVIDNet* (w/o Transformer)" did not include the Transformer in the country-level encoder (see Figure 2). The RMSE was again increased by up to 12.3%. Therefore, we confidently confirm that *Hi-COVIDNet* contains essential components to understand the epidemic risk.

5 RELATED WORK

Owing to the urgency in ending the COVID-19 pandemic, numerous studies have tried to predict the spread of COVID-19 by applying various methods developed in the infectious disease and pandemic research community. Recently, machine learning and deep learning are actively exploited to enjoy their powerful performance proven in other fields [15, 17].

5.1 Machine Learning Approach

As a traditionally important problem, Yang et al. [29] applied the susceptible exposed infected resistant (SEIR) model to estimate the epidemic curve of COVID-19. They utilized the data of severe acute respiratory syndrome (SARS) since 2003 to train a model and applied it to predict the COVID-19 transmissions. For the same objective, the ARIMA model [1] and mathematical modeling [12] were applied to figure out the degree of COVID-19 transmissions.

Several studies employed popular machine learning techniques instead of traditional statistical analysis. Hamer et al. [7] evaluated statistical measures using a decision tree and a boosted decision tree to predict the spatio-temporal epidemic spread and its risk. Machando et al. [16] applied random forests [2] and other several algorithms with the animal movements and spatial contexts to expect the outbreak of porcine epidemic diarrhea virus (PEDV), which is a kind of coronaviruses but only infectious to animals. Li et

al. [14] estimated the number of confirmed cases around the world with simple linear regression using the data published by WHO.

5.2 Recent Deep Learning Approach

The great success of deep learning in various applications also attracted pandemic related research. Pal et al. [20] and Uhlig et al. [26] exploited an artificial neural network to predict the spread of COVID-19. Chimmula et al. [3] predicted the number of confirmed cases and the end date of COVID-19 in Canada, Italy, and USA using a LSTM. Punn et al. [23] estimated the numbers of diagnosed, dead, and released cases and compared the accuracy obtained by various models such as a DNN, a RNN, and a LSTM.

For infectious diseases other than COVID-19, Wu et al. [28] used the RNN and CNN for epidemiological predictions; the RNN is used to capture the long-term temporal correlation in the data, and the CNN is used to fuse the information from different sources. Mussumeci et al. [19], using multivariate time-series as predictors, investigated the spatial effects on the spread of Dengue fever, and the LSTM-based model was shown to achieve the lowest error rate. For the interested reader, please refer to the survey article by Philemon et al. [22] that reviews the neural network components used for infectious disease prediction.

While previous studies mostly focused on predicting the epidemic trend itself, to the best of our knowledge, our work is the first attempt to predict the trend of "imported cases," which is really useful for the quarantine *at the border*.

6 CONCLUSION

In this paper, we proposed a novel approach *Hi-COVIDNet* to address the problem of predicting imported COVID-19 cases, which is an urgent and significant issue to control the disease. *Hi-COVIDNet* understands the temporal dependency of COVID-19 infections country-wise as well as the interaction from each country, followed by the incorporation of the geographic hierarchy per continent. We showcased the practicality and effectiveness of *Hi-COVIDNet* through a case study in South Korea. *Hi-COVIDNet* predicted the upcoming number of imported COVID-19 cases much more precisely than the baselines. Overall, we believe that our work can assist the governments in adjusting their special entry procedure to a new situation rapidly. We are working closely with the Korean government to put our work to practical use and are very interested in applying *Hi-COVIDNet* to other countries.

ACKNOWLEDGEMENT

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00862, DB4DL: High-Usability and Performance In-Memory Distributed DBMS for Deep Learning). We appreciate KT for providing the roaming dataset.

REFERENCES

[1] Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., and Ciccozzi, M. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief* (2020), 105340.
[2] Breiman, L. Random forests. *Machine Learning 45*, 1 (2001), 5–32.
[3] Chimmula, V. K. R., and Zhang, L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals* (2020), 109864.
[4] Contreras, J., Espinola, R., Nogales, F. J., and Conejo, A. J. ARIMA models to predict next-day electricity prices. *IEEE Transactions on Power Systems 18*, 3 (2003), 1014–1020.
[5] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255.
[6] Domany, E., van Hemmen, J. L., and Schulten, K. *Models of neural networks III: Association, generalization, and representation.* Springer, 1996.
[7] Hamer, W. B., Birr, T., Verreet, J.-A., Duttmann, R., and Klink, H. Spatio-temporal prediction of the epidemic spread of dangerous pathogens using machine learning methods. *ISPRS International Journal of Geo-Information 9*, 1 (2020), 44.
[8] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural Computation 9*, 8 (1997), 1735–1780.
[9] Ioffe, S., and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
[10] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[11] Krogh, A., and Hertz, J. A. A simple weight decay can improve generalization. In *Proceedings of Advances in Neural Information Processing Systems* (1992), pp. 950–957.
[12] Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., Sun, F., Jit, M., Munday, J. D., et al. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *The Lancet Infectious Diseases* (2020).
[13] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.
[14] Li, Y., Liang, M., Yin, X., Liu, X., Hao, M., Hu, Z., Wang, Y., and Jin, L. COVID-19 epidemic outside China: 34 founders and exponential growth. *medRxiv* (2020).
[15] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis 42* (2017), 60–88.
[16] Machado, G., Vilalta, C., Recamonde-Mendoza, M., Corzo, C., Torremorell, M., Perez, A., and VanderWaal, K. Identifying outbreaks of Porcine Epidemic Diarrhea virus through animal movements and spatial neighborhoods. *Scientific Reports 9*, 1 (2019), 1–12.
[17] McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730* (2018).
[18] McNeil, D. G. Coronavirus has become a pandemic, W.H.O. says. https://www.nytimes.com/2020/03/11/health/coronavirus-pandemic-who.html.
[19] Mussumeci, E., and Coelho, F. C. Machine-learning forecasting for Dengue epidemics - Comparing LSTM, random forest and lasso regression. *medRxiv* (2020).
[20] Pal, R., Sekh, A. A., Kar, S., and Prasad, D. K. Neural network based country wise risk prediction of COVID-19. *arXiv preprint arXiv:2004.00959* (2020).
[21] Park, S. N. Cults and conservatives spread coronavirus in South Korea. https://foreignpolicy.com/2020/02/27/coronavirus-south-korea-cults-conservatives-china/.
[22] Philemon, M. D., Ismail, Z., and Dare, J. A review of epidemic forecasting using artificial neural networks. *International Journal of Epidemiologic Research 6*, 3 (2019), 132–143.
[23] Punn, N. S., Sonbhadra, S. K., and Agarwal, S. COVID-19 epidemic analysis using machine learning and deep learning algorithms. *medRxiv* (2020).
[24] Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Proceedings of Advances in Neural Information Processing Systems* (2019), pp. 1917–1928.
[25] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*, 1 (2014), 1929–1958.
[26] Uhlig, S., Nichani, K., Uhlig, C., and Simon, K. Modeling projections for COVID-19 pandemic by combining epidemiological, statistical, and neural network approaches. *medRxiv* (2020).
[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems* (2017), pp. 5998–6008.
[28] Wu, Y., Yang, Y., Nishiura, H., and Saitoh, M. Deep learning for epidemiological predictions. In *Proceedings of 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (2018), pp. 1085–1088.
[29] Yang, Z., Zeng, Z., Wang, K., Wong, S.-S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease 12*, 3 (2020), 165.
[30] Zhai, P., Ding, Y., Wu, X., Long, J., Zhong, Y., and Li, Y. The epidemiology, diagnosis and treatment of COVID-19. *International Journal of Antimicrobial Agents* (2020), 105955.

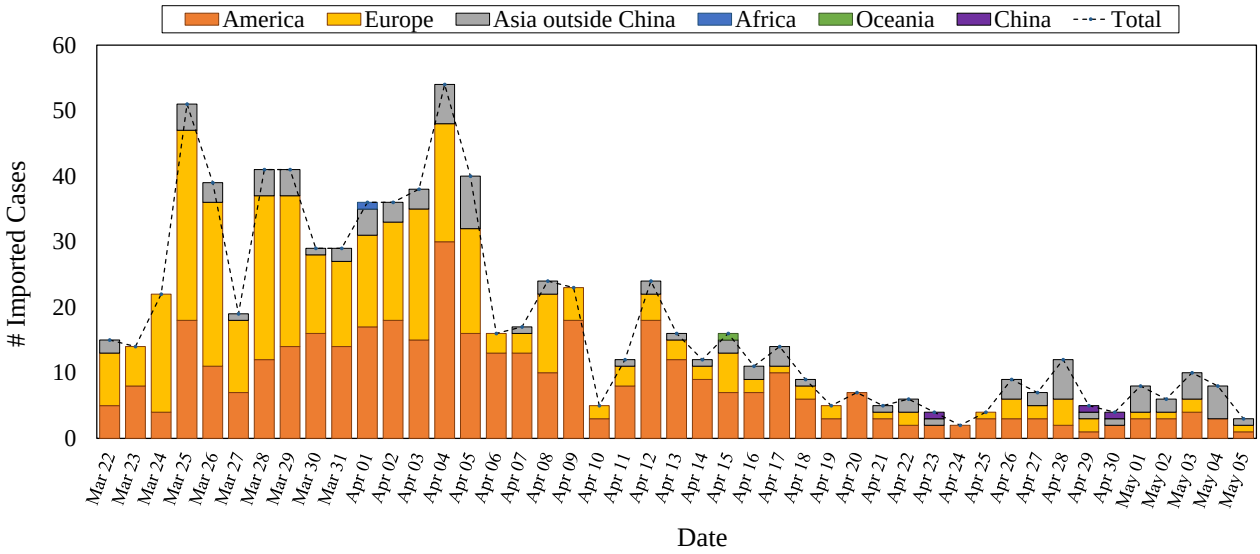


Figure 5: The trend of the imported COVID-19 cases in South Korea during the training period (best viewed in color). The imported cases are counted for each continent.

A APPENDIX

A.1 Imported Cases in the Training Period

For the interested reader, in Figure 5, we show the statistics of the imported cases to South Korea for the training period from March 22 through May 5, 2020, which are provided by KCDC. Most imported cases were from America and Europe, especially during the early days (i.e., late March and early April) when COVID-19 started spreading rapidly all around the world. On the other hand, there were consistently few imported cases from Oceania, China, and Africa during the entire training period. The total number of imported cases decreased after early April, because the number of inbound passengers from America and Europe suddenly dropped by the COVID-19 lockdown policy of the countries. Notably, capturing this dynamic trend within a short period is very challenging, and *Hi-COVIDNet* successfully demonstrated a potential for the task.