Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 3, pp. 272-279, June 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.3.272

<Application Research>

# 머신러닝 모델을 이용한 대한민국 코로나 신규 확진자 예측

배진수 · 김성범†

고려대학교 산업경영공학과

# Predictions of COVID-19 in Korea Using Machine Learning Models

Jinsoo Bae · Seoung Bum Kim

School of Industrial and Management Engineering, Korea University

COVID-19, a coronavirus (COVID-19) caused 57,680 confirmed cases and 819 deaths in Korea as of December 28, 2020, causing many casualties worldwide. Predicting COVID-19 confirmed cases allows us to manage and plan effective preventive measures to reduce casualties. In this paper, we propose a methodology to predict COVID-19 new confirmed cases over the next four days using machine learning models. We propose using long short-term memory (LSTM), random forest, gradient boosting models. Experiments show that LSTM produces better prediction performance over other models in the majority of scenarios. We believe that this study is the first attempt to predict the trend of COVID-19 confirmed cases in Korea. We hope our work can inspire researchers to develop better methods to predict COVID-19 confirmed cases.

***Keywords:*** COVID-19, Forecasting, LSTM, Random Forest, Gradient Boosting

## 1. 서 론

코로나바이러스감염증-19, 이하 COVID-19는 새로운 유형의 코로나바이러스(SARS-CoV-2)에 의한 호흡기 감염질환으로 감염자의 비말(침방울)이 호흡기나 눈, 코, 입의 점막으로 침투될 때 전염되며 강한 전파력을 갖고 있다. 2019년 12월 중국 우한에서 처음 발생한 이후 전 세계로 확산되며 많은 인명 피해를 끼치고 있으며(Zu *et al*., 2020), 국내에는 2020년 12월 28일 기준 57,680명 확진자와 사망자 819명이 발생하였다(Ministry of Health and Welfare Organization. 2020). <Figure 1>은 2020년 1월 20일 국내 첫 확진자가 발생한 이후 2020년 12월 25일까지의 국내에 발생한 코로나 신규 확진자 발생 패턴을 보여준다. 2020년 2월 18일 대구 신천지 교회와 8월 13일 서울 사랑제일교회에서 발생한 대규모 집단 감염부터 여러 차례의 소규모 지역 감염을 계기로 대한민국 전체에 코로나 확진자가 퍼지게 되었다. 또한, 11월 10일 수도권을 중심으로 3차 코로나바이러스 대유행이 시작되었으며, 하루 1000명에 가까운 확진자가 발생하는 것을 확인할 수 있다. 이처럼 코로나는 강한 전파력을 가졌기 때문에 많은 인명 피해를 일으키고 있으며, 현재까지 확실한 백신이 도입되지 않은 상태이다(Grenfell and Drew, 2020). 코로나 전파를 막기 위한 유일한 방법은 철저한 개인 위생 관리와 강도 높은 사회적 거리두기 운동이며, 현재 꾸준하게 실천되고 있음에도 확진자가 계속 발생하고 있다.

전 세계적으로 코로나 바이러스로 인해 발생할 인명 피해를 줄이고자 코로나바이러스 관련 데이터 분석을 다양하게 수행하고 있다. He *et al*.(2020)는 Susceptible-exposed-infected-removed 질병 확산 모형과 Particle swarm optimization 알고리즘을 통해 중국 후베이의 코로나 확산 현상을 시뮬레이션을 통해 구현하였다. Alazab *et al*.(2020)는 오스트레일리아와 요르단 국가에 발생한 코로나 확진자, 회복자, 사망자 수 예측에 Prophet 모델이 효과적임을 보였다. Arora *et al*.(2020)는 인도에서 발생한 코로나 확진자 추세를 순환신경망 모델로 예측하는 연구를 실시하였다. Chimmula and Zhang(2020)는 순환신경망 모델로 캐나다에서 발생한 코로나 확진자 추세와 종식 시기를 예측하는 연구를 진행하였다. 하지만 선행 연구들은 코로나 확진자 변수만을 사용한 단변량 회귀 모델을 사용하였고, 확진자가 급변하는

† 연락저자 : 김성범 교수, 02841 서울특별시 성북구 안암로 145 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-3290-4550, E-mail : sbkim1@korea.ac.kr
2021년 1월 4일 접수; 2021년 2월 16일 수정본 접수; 2021년 2월 25일 게재 확정.

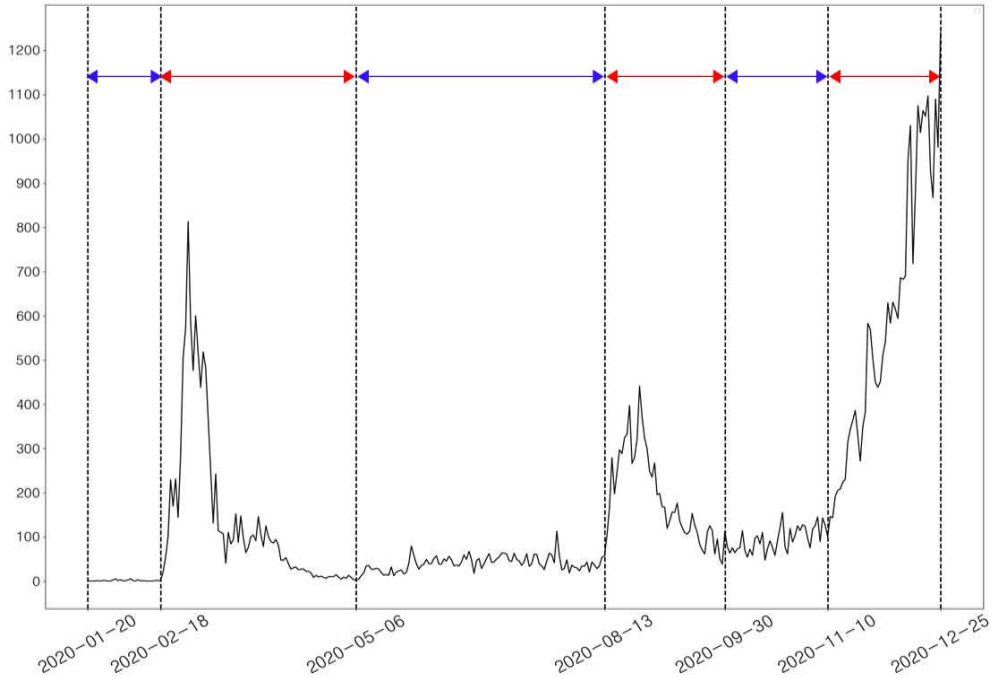



**Figure 1.** COVID-19 Confirmation Trend in Korea. The Black Solid Line Indicates Number of New COVID-19 Confirmed Cases by Date. The Red Line Indicates Period when Trend of COVID-19 Confirmed Cases Changes Rapidly Due to Group Infection. The Blue Line Indicates Period when COVID-19 Infections Occur Sporadically

구간과 불규칙적인 구간을 구분하지 않고 예측 성능을 평가하였기 때문에, 확진자 변동의 특징에 적합한 예측 모델을 제안하지 못하였다. 이를 보완하여 대한민국 코로나 확진자에 대한 체계적인 예측이 가능하다면 인명 피해를 줄이는 데 효과적인 대비책을 세울 수 있게 될 것이다.

본 논문은 머신러닝을 이용한 국내 코로나 신규 확진자 추세 예측 방법론을 제안한다. 과거에 발생한 코로나 확진자 추세와 법적 공휴일 여부를 기반하여 4일 뒤 코로나 신규 확진자를 예측하는 자기 회귀 방법론으로 장기-단기 기억 신경망(long short-term memory model, LSTM), 랜덤 포레스트(random forest), 그레디언트 부스팅(gradient boosting) 모델의 사용을 제안하였다. LSTM은 시계열 데이터 예측에 정확도가 높은 순환신경망 계열의 딥러닝 모델로 다른 순환 신경망 모델보다 단순한 구조를 가져 적은 데이터 개수로도 학습이 가능하다. 또한 메모리 셀(memory cell)로 장기 의존성 문제를 해결하여 안정적인 학습이 가능하고 시계열 데이터 학습에 효과적인 모델이다. 랜덤 포레스트, 그레디언트 부스팅은 여러 개의 모델을 결합해 하나의 모델보다 더 우수한 예측 성능을 갖는 앙상블(ensemble) 계열의 모델로, 코로나 확진자 추세와 같이 비선형 구조를 가진 데이터 분석에 효과적이다. 본 연구는 위 세가지 모델이 코로나 신규 확진자 추세 예측에 효과적임을 입증하기 위해, 여러 가지 시나리오에 대한 실험을 진행하였고, 실험 결과 LSTM 모델이 전반적으로 우수한 예측 성능을 보였다. 특히 코로나 신규 확진자가 대규모 집단 감염으로 인해 급증하고 감소하는 추세를 올바르게 예측하였고, 그레디언트 부스팅과 랜덤 포레스트는 3차 대유행의 초기패턴과 산발적으로 발생한 소규모 지역 감염의 추세를 올바르게 예측하였다. 본 논문은 국내 코로나 확진자 추세 예측에 머신러닝을 체계적으로 시도한 첫 번째 연구이며, 향후에 일어날 코로나 확진자 증가에 대한 예방책 마련과 추가 연구에 대해 도움될 수 있을 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제안 방법론에 대한 자세한 설명을 제 2장에서 실시하고, 제 3장에서 제안 방법론에 의해 예측된 국내 코로나 신규 확진자 결과를 소개한다. 마지막으로 제 4장에서 결론 및 향후 연구에 대해 다루도록 한다.

## 2. 제안 방법론

### 2.1 데이터 수집 및 전처리

본 논문에서 사용한 원본 데이터는 질병관리청이 운영하는 코로나바이러스감염증-19 정식 보도자료(http://ncov.mohw.go.kr/)에서 획득하였고 수집 기간과 데이터 내용은 <Table 1>에서 확인할 수 있다. <Table 1>의 변수 1부터 변수 6은 각 일별 신규 확진자와 1일부터 5일 전 신규 확진자 수를 의미하고, 변수 7부터 변수 12는 당일부터 5일 전까지의 법적 공휴일 여부를 나타내는 이진변수(binary variable)이다. 수집된 데이터의 12가지 변수는 최소최대변환(minmax scaling)으로 식 (1)과 같이 변환하였고, 일주일 단위로 종합(식 (2))하는 전처리 과정을 실시하였다. 전처리가 완료된 데이터는 총 84개의 입력변수와 종속변수인 4일 후 코로나 신규 확진자 수로 구성되었으며 식 (3)과 같다.

**Table 1.** Description of Raw Collected Data

| Column index | Contents (collection period : 2020. 01. 14~2020. 12. 25) |
|---|---|
| 1 | Number of new Corona confirmed today |
| 2 | Number of new Corona confirmed one days ago |
| 3 | Number of new Corona confirmed two days ago |
| 4 | Number of new Corona confirmed three days ago |
| 5 | Number of new Corona confirmed four days ago |
| 6 | Number of new Corona confirmed five days ago |
| 7 | Whether or not, today is a legal holiday |
| 8 | Whether or not, one days ago is a legal holiday |
| 9 | Whether or not, two days ago is a legal holiday |
| 10 | Whether or not, three days ago is a legal holiday |
| 11 | Whether or not, four days ago is a legal holiday |
| 12 | Whether or not, five days ago is a legal holiday |

### 2.2 예측 방법

본 연구는 당일부터 5일전까지 코로나 확진자 수와 법적 공휴일 여부인 $x_i^{scaled}$(식 (2))를 일주일간 종합한 $X_i$(식 (3))로 4일 후 코로나 신규 확진자인 $y_i$(식 (3))를 예측하기 위해 LSTM, 랜덤 포레스트, 그레디언트 부스팅 모델을 사용하였다. 예측에 대한 전반적인 프로세스는 <Figure 2>에서 확인할 수 있다. 먼저 수집된 원본 데이터에 전처리(식 (1)~식 (3))을 수행한 뒤, 훈련용 데이터(training data)와 평가용 데이터(testing data)로 분리한다. 그리고 훈련용 데이터로 예측 모델을 학습시킨 뒤, 학습된 모델에 평가용 데이터를 입력해 4일 후 코로나 신규 확진자를 예측한다.

$$x_{ik}^{scaled} = \frac{x_{ik} - \min_i x_{ik}}{\{(\max_i x_{ik}) - (\min_i x_{ik})\}} \quad for\ all\ i, k \in A, \qquad (1)$$

$$where\ A = \{(i, k) \mid i = 1, 2, 3, \cdots 347\ and\ k = 1, 2, 3, \cdots, 12\}$$

$$X_i = concat(x_i^{scaled}, x_{i-1}^{scaled}, x_{i-2}^{scaled}, x_{i-3}^{scaled}, x_{i-4}^{scaled}, x_{i-5}^{scaled}, x_{i-6}^{scaled}) \qquad (2)$$

$$where\ x_i^{scaled} = [x_{i1}^{scaled}, x_{i2}^{scaled}, x_{i3}^{scaled}, \cdots, x_{i12}^{scaled}]^T$$

$$Preprocessed\ dataset = \{(X_i, y_i) \mid y_i\ is\ the\ number\ of\ COVID\text{-}19\ confirmed\ cases\ of\ i+4th\ date\} \qquad (3)$$

LSTM은 순차적으로 등장하는 시계열 데이터를 학습하는데 효과적인 순환신경망(recurrent neural network, RNN) 계열의 모델로 일반적인 인공신경망과 달리 내부 상태(internal state)를 통해 입력 받은 정보들의 특징을 기억할 수 있다(Hochreiter and Schmidhuber, 1997). LSTM은 RNN이 이전 시점의 정보를 오랫동안 기억하지 못하고 그레디언트가 소멸하는 문제(장기 의존성 문제)를 해결하여 안정적인 학습이 가능한 모델로, 이전 시점의 정보를 효과적으로 기억하는 메모리 셀(memory cell)로 장기 의존성 문제를 해결하였다. Memory cell은 현재 시점 정보 중 memory cell($C_t$)에 담고자 하는 정보를 제어하는 입력 게이트(식 (4)), 이전 시점 memory cell($C_{t-1}$) 내에 불필요한 정보를 지우는 망각 게이트(식 (5)), memory cell을 이용해 hidden state를 출력하는 출력 게이트(식 (6))로 작동되며 세 가지 게이트가 연산하는 수식과 LSTM cell 구조는 <Figure 3>과 같다.

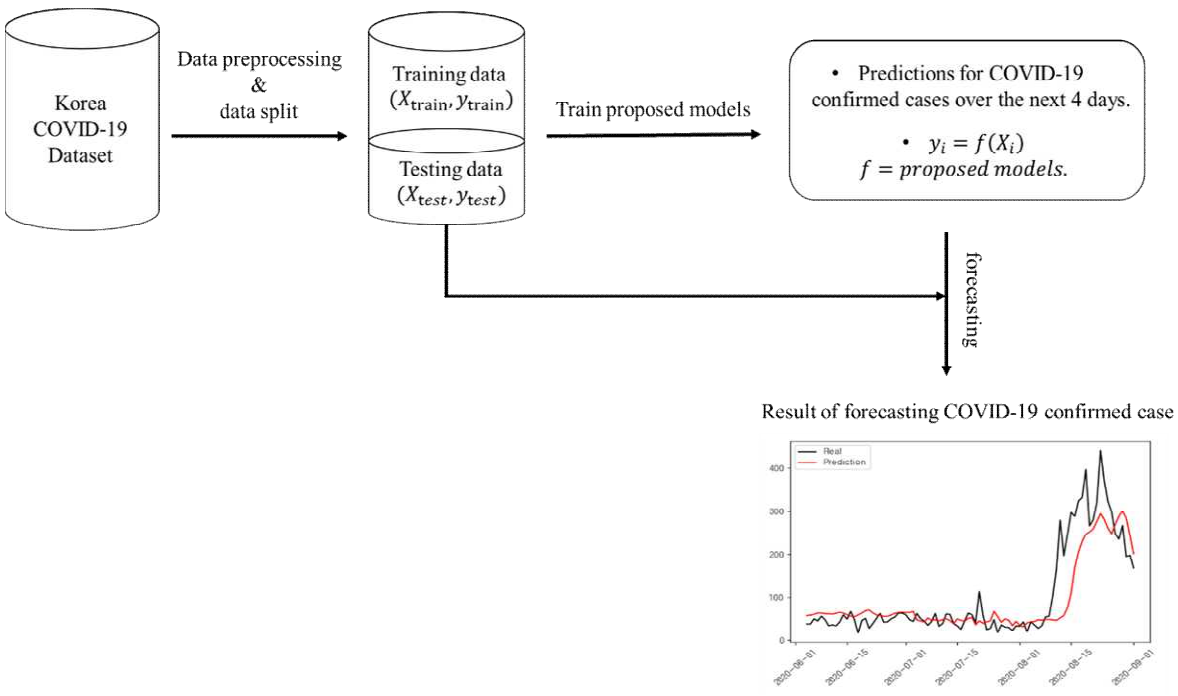
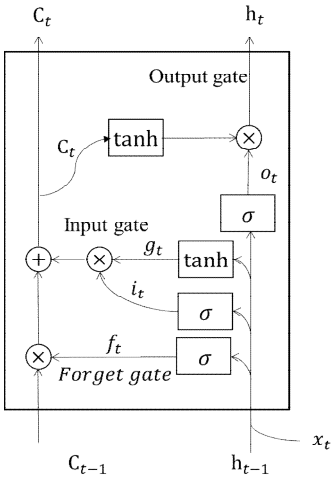



**Figure 2.** Overall Process of the Proposed Method




$$i_t = \sigma(W_{xh_i}x_t + W_{hh_i}h_{t-1} + b_i) \quad (4)$$
$$g_t = \tanh(W_{xh_g}x_t + W_{hh_g}h_{t-1} + b_g)$$

$$f_t = \sigma(W_{xh_f}x_t + W_{hh_f}h_{t-1} + b_f) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$o_t = \sigma(W_{xh_o}x_t + W_{hh_o}h_{t-1} + b_o) \quad (6)$$
$$h_t = o_t \odot \tanh(c_t)$$

**Figure 3.** Structure of an LSTM Cell and Equations that Describe the Gates of an LSTM Cell

식 (4), 식 (5)에서 $W_{xh_i}$, $W_{xh_g}$, $W_{xh_f}$, $W_{hh_i}$, $W_{hh_g}$, $W_{hh_f}$와 $b_i$, $b_g$, $b_f$는 입력 게이트와 망각 게이트에 대한 파라미터와 바이어스를 의미하고, 식 (6)에서 $W_{xh_o}$, $W_{hh_o}$와 $b_o$는 출력 게이트에 대한 파라미터와 바이어스를 나타낸다. $t-1$와 $t$은 이전 시점과 현재 시점을 의미하고 LSTM cell은 입력 벡터 $x_t$, 히든 벡터 $h_{t-1}$,메모리 셀 $C_{t-1}$을 입력 받아 메모리 셀 $C_t$과 히든 벡터 $h_t$를 출력한다 (식 (4)~식 (6)). 이 때, 입력 벡터 $x_1$, $x_2$, $x_3$, …, $x_7$은 식 (2)의 $x_i^{scaled}$, $x_{i-1}^{scaled}$, $x_{i-2}^{scaled}$, …, $x_{i-12}^{scaled}$,를 의미하고, 히든 벡터 $h_7$를 레이어가 2개이고 정류된 선형 유닛 함수(rectified linear unit)를 활성화함수로 가진 순방향 인공신경망(feed forward neural network)에 입력(식 (7))하여 4일 뒤 신규 확진자 $y_i$를 예측한다.본 논문에서 사용한 LSTM 모델의 하이퍼파라미터는 <Table 2>와 같다.

$$y_i = W_b(Relu(W_1h_7 + b_1)) + b_2 \quad (7)$$

랜덤 포레스트는 여러 개의 의사결정나무(decision tree)를 결합한 모델로 하나의 의사결정나무보다 더 우수한 예측 성능을 갖는다(Pavlov, 2019). 의사결정나무는 계층적 구조의 여러 의사 결정 규칙들을 이용해 예측을 실시하는 모델이며(Rathore and Kumar, 2016), 랜덤 포레스트가 여러 개의 의사결정나무 모델로 결합된 것의 장점은 일부 의사결정나무가 잘못된 예측을 진행하더라도 여러 의사결정나무 모델의 예측 결과를 종합함으로써 정확한 예측이 가능한 것에 있다. 이는 모델의 일반화 성능을 보장할 수 있음을 의미하며 랜덤 포레스트의 구조는

**Table 2.** Description of Hyperparameter of LSTM which Used in this Paper

| Hyperparameter | Description | Values |
|---|---|---|
| Learning rate | Learning rate for LSTM parameter updates | 0.01 |
| Optimizer | The name of optimizer for reducing loss function | Adadelta |
| Batch size | The number of training examples in one forward/backward pass | 10 |
| Loss function | A criterion that measures between predicted values and real values | Mean squared error |
| Hidden size | The number of features in the hidden vector $h_t$ | 12 |
| Number of neurons | The $i$th element represents the number of neurons in the $i$th hidden layer of feed forward neural network | (12, 20) |

**Table 3.** Description of Hyperparameter of Random Forest which used in this Paper

| Hyperparameter | Description | Values |
|---|---|---|
| Number of trees | Number of trees in the random forest | [50, 100, 300] |
| Max depth | The maximum depth of the tree. | [3, 10, 15, 20, None] |
| Splitting rules | Splitting criteria in the nodes | Mean squared error |

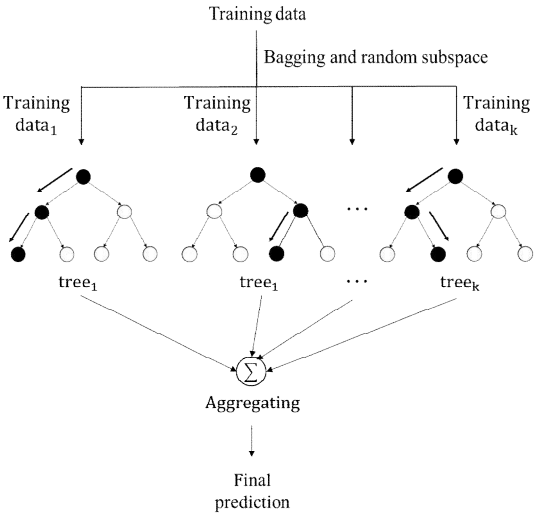



**Figure 4.** Overview of a Random Forest Model

<Figure 4>와 같고 탐색한 하이퍼파라미터는 <Table 3>과 같다.

그레디언트 부스팅은 여러 개의 의사결정나무의 예측 결과들을 종합해 예측한다는 점에서는 랜덤 포레스트와 동일하지만, 의사결정나무를 순차적으로 학습시킨다는 점이 가장 큰 특징이다(Friedman, 2002). 이전 의사결정나무가 잘못 예측하였던 에러를 다음 의사결정나무가 학습하여 점진적으로 에러를 줄여가는 순차적인 학습을 진행하며, 모델 구조와 탐색한 하이퍼파라미터는 <Figure 5>와 <Table 4>와 같다.

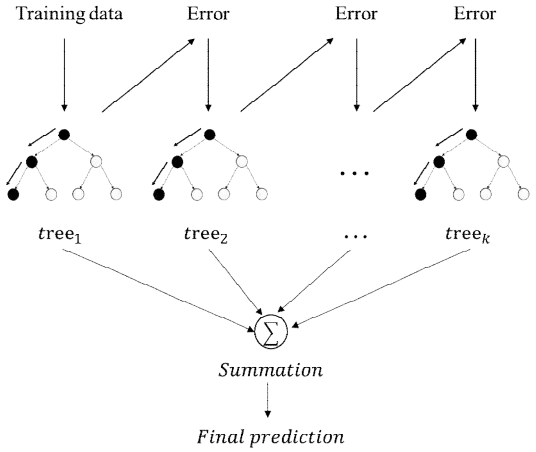


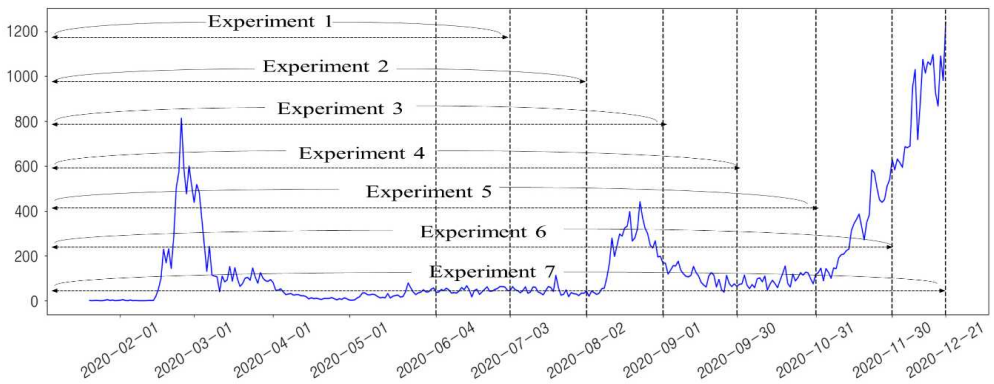

**Figure 5.** Overview of a Gradient Boosting Model

**Table 4.** Description of Hyperparameter of Gradient Boosting Model which Used in this Paper

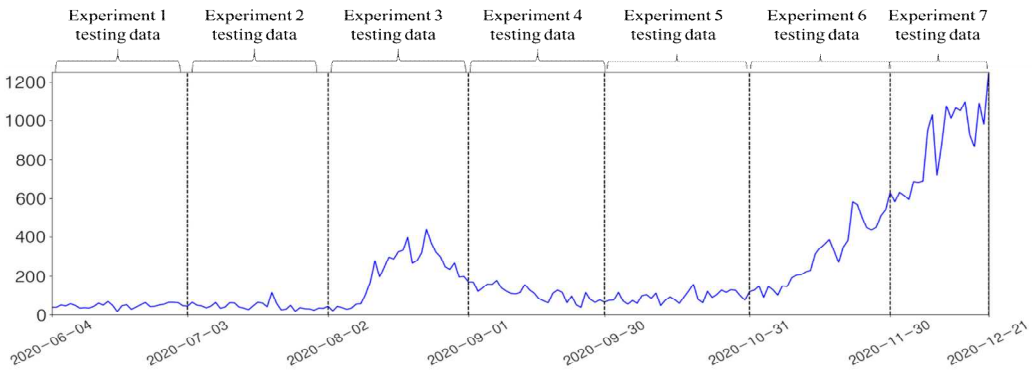
| Hyperparameter | Description | Values |
|---|---|---|
| Number of trees | Number of trees in the random forest | [50, 100, 300] |
| Max depth | The maximum depth of the tree. | 3 |
| Splitting rules | Splitting criteria in the nodes | [Friedman mean squared error, Mean absolute error] |

## 3. 예측 결과

4일 후 신규 확진자 수 예측 모델 성능을 시점별로 파악하고자 <Table 5>와 같이 7가지 실험 기간을 설정하였고, 각 실험 기간별 데이터 개수는 실험 기간 일수와 동일하다. 실험 1과 실험 2, 실험 5의 예측 기간은 지역 감염으로 인한 신규 확진자 발생이 산발적으로 지속되는 시기였으며, 실험 3과 실험 4의 예측 기간은 국내 종교 집단 포함 감염으로 확진자가 급증하는 시기와 감소하는 시기를 포함한 기간이며, 실험 6과 실험 7은 수도권을 중심으로 확진자가 급증하는 3차 대유행 기간이다. 각 실험 별 훈련 및 검증용 데이터 기간의 확진자 추세는 <Figure 6(a)>에서 확인할 수 있으며, 평가용 데이터의 확진자 발생 추세는 <Figure 6(b)>에서 볼 수 있다. 일곱가지 실험에 대한 예측 결과를 분석하기 위해 사용한 지표는 평균 제곱근 오차(root mean square error, RMSE)와 평균 제곱근 로그 오차(root mean squared logarithmic error, RMSLE)이며 식 (8)~식 (9)와 같이 계산된다. 평균 제곱근 오차는 확진자 수와 동일한 단위를 가져 직관적으로 예측 성능을 평가할 수 있으며, 각 실험별 최적 모델을 선택하는데 사용하였다. 평균 제곱근 로그 오차는 상대적 크기에 기반하여 오차를 산출하기 때문에 확진자 변동이 큰 시기에도 지표의 변화가 크지 않으며 과다예측보다 과소예측 시 더 큰 에러를 부여하기 때문에 각 실험별 최적 모델 선택에 실용성을 함께 고려할 수 있다.




(a)




(b)

**Figure 6.** (a) Training/validation data, and (b) testing data for seven experimental scenarios

**Table 5.** Period and Number of Data for Seven Experimental Scenarios

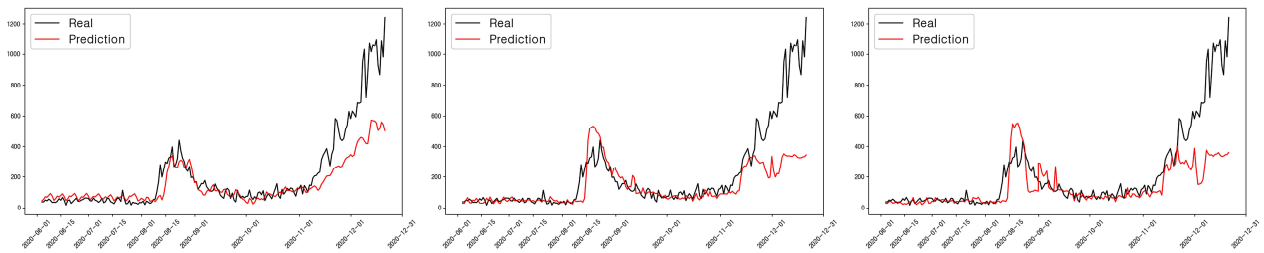
| | Experiment 1 | Experiment 2 | Experiment 3 | Experiment 4 | Experiment 5 | Experiment 6 | Experiment 7 |
|---|---|---|---|---|---|---|---|
| Training data | 2020.01.20. ~2020.05.24. (126) | 2020.01.20. ~2020.06.23. (156) | 2020.01.20. ~2020.07.23. (186) | 2020.01.20. ~2020.08.21. (215) | 2020.01.20. ~2020.09.20. (245) | 2020.01.20. ~2020.10.21. (276) | 2020.01.20. ~2020.11.20. (306) |
| Validation data | 2020.05.25. ~2020.06.03. (10) | 2020.06.24. ~2020.07.03. (10) | 2020.07.24. ~2020.08.02. (10) | 2020.08.22. ~2020.08.31. (10) | 2020.09.21. ~2020.09.30. (10) | 2020.10.22. ~2020.10.31. (10) | 2020.11.21. ~2020.11.30. (10) |
| Testing data | 2020.06.04. ~2020.07.03. (30) | 2020.07.04. ~2020.08.02. (30) | 2020.08.03. ~2020.09.01. (30) | 2020.09.01. ~2020.09.30. (30) | 2020.10.01. ~2020.10.31. (31) | 2020.11.01. ~2020.11.30. (30) | 2020.12.01. ~2020.12.21. (21) |

**Table 6.** RMSE of LSTM, Random Forest, Gradient Boosting

| | Experiment 1 | Experiment 2 | Experiment 3 | Experiment 4 | Experiment 5 | Experiment 6 | Experiment 7 | Average |
|---|---|---|---|---|---|---|---|---|
| LSTM | 25.68 | 30.65 | 76.61 | 35.75 | 32.84 | 155.60 | 433.72 | 112.98 |
| random forest | 18.99 | 18.19 | 119.12 | 44.86 | 34.48 | 158.58 | 593.86 | 141.15 |
| gradient boosting | 19.51 | 19.45 | 138.23 | 57.35 | 32.86 | 147.34 | 594.85 | 144.23 |

**Table 7.** RMSLE of LSTM, Random Forest, Gradient Boosting

| | Experiment 1 | Experiment 2 | Experiment 3 | Experiment 4 | Experiment 5 | Experiment 6 | Experiment 7 | Average |
|---|---|---|---|---|---|---|---|---|
| LSTM | 0.212 | 0.353 | 0.27 | 0.135 | 0.194 | 0.247 | 0.393 | 0.258 |
| random forest | 0.19 | 0.14 | 0.441 | 0.14 | 0.155 | 0.273 | 1.084 | 0.346 |
| gradient boosting | 0.233 | 0.164 | 0.57 | 0.161 | 0.132 | 0.239 | 1.195 | 0.385 |




(a) LSTM (b) random forest (c) gradient boosting

**Figure 7.** Predicted COVID-19 Confirmed Cases by LSTM, Random Forest, Gradient Boosting. The Black and Red Lines Indicate the Actual and Predicted Cases, Respectively

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_l)^2} \quad (8)$$

$$RMSLE = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(\log(\hat{y}_l + 1) - \log(y_i + 1))\right)^2} \quad (9)$$

where $y_i$ represents COVID-19 confirmed in 4 days and $\hat{y}_l$ represents predicted values.

각 실험 시나리오별 RMSE와 RMSLE는 <Table 6>과 <Table 7>에서 확인할 수 있고, 세 가지 모델에 의해 예측된 확진자 추세는 <Figure 7>에서 확인할 수 있다. 여러 시나리오에 대한 평균 예측 성능 기준 LSTM이 가장 우수한 성능을 보였고, 랜덤포레스트와 그레디언트 부스팅은 비슷한 예측 성능을 보였다. RMSE와 RMSLE를 기준으로 실험1과 2는 랜덤 포레스트 모델이 가장 우수한 예측 성능을 보였으며, 산발적으로 발생한 소규모 지역 감염 예측에 랜덤 포레스트가 정확함을 알 수 있었다. 확진자가 처음으로 폭증하는 구간을 담고 있는 실험 3과 확진자가 감소하는 추세를 보인 실험 4에서는 RMSE와 RMSLE를 기준으로 LSTM 모델이 다른 모델보다 전반적으로 우수한 예측력을 보였음을 <Figure 7(a)>에서 확인할 수 있다. 소규모로 발생한 지역 감염이 지속되는 시기를 포함한 실험 5는 RMSE를 기준으로 예측 성능 차이가 1~2명으로 크지 않았으나, RMSLE를 통해 그레디언트 부스팅이 효과적임을 확인할 수 있었다. 3차 대유행 기간의 시작 구간을 포함하고 있는 실험 6에서는 그레디언트 부스팅 모델이 유행의 시작구간을 올바르게 예측하였음을 <Figure 7(c)>에서 확인할 수 있다. 3차 대유행의 폭증 추세가 포함된 실험 7에서는 LSTM 모델이 가장 우수한 예측 성능을 보여주었다.

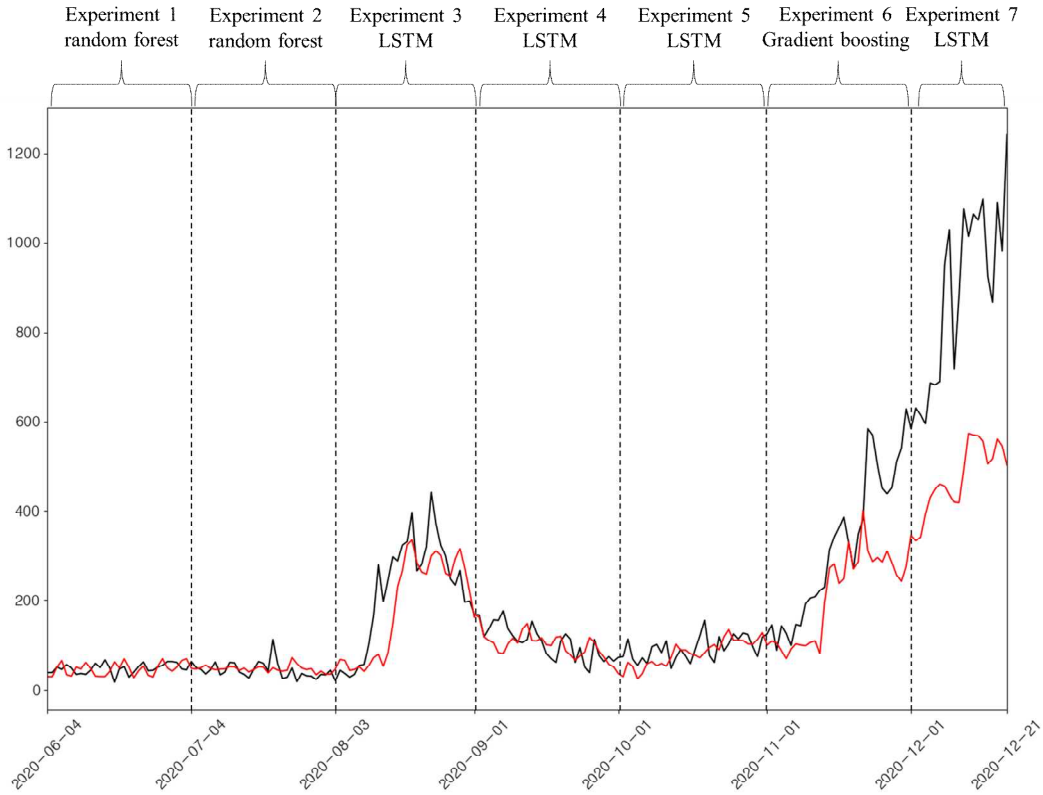



**Figure 8.** Predicted COVID-19 Confirmed Cases by the Best Model in Each Experimental Period. The Black and Red Lines Indicate the Actual and Predicted Cases, Respectively

각 실험 시나리오별 최적 RMSE를 가진 모델의 예측 결과를 종합해보면 <Figure 8>과 같다. 예측 결과 머신러닝 모델이 미세한 변동 패턴과 피크 형태의 패턴은 정확하게 예측하지 못하였지만 전반적인 변동 패턴은 올바르게 예측하고 있음으로 볼 수 있었다.

## 4. 결 론

코로나바이러스는 전세계적으로 퍼져 많은 인명 피해와 경제적 손실을 가져왔다. 현재까지 확실한 백신과 치료제가 도입되지 않았으며 철저한 개인 위생 관리와 강도 높은 사회적 거리두기 운동이 진행되고 있음에도 불구하고 바이러스 전파가 지속적으로 발생하고 있다. 해외에는 머신러닝 기술을 이용하여 향후 코로나 신규 확진자를 예측하는 연구가 진행되었으나, 코로나 확진자 정보만을 학습한 예측 모델을 사용하였고 짧은 예측 성능 평가 기간과 체계적인 실험을 진행하지 않아 효과적인 코로나 확진자 수 예측 모델을 올바르게 제안하였다고 보기 어렵다. 본 연구에서는 과거로부터 현재까지 대한민국에서 발생한 코로나 확진자 정보와 법적 공휴일 여부를 이용하여 4일 후 신규 확진자를 예측하는 방법론을 제안하였다. 한달 간격의 예측 기간을 7개 두어 예측 모델의 성능을 평가하였으며, 실험 결과 확진자가 급증하는 초기 패턴과 소규모의 불규칙적인 지역 감염 예측에는 앙상블 계열의 모델이 효과적이며, 확진자가 급증하거나 감소하는 시기에 대한 예측은 LSTM 모델이 효과적임을 확인할 수 있었다.

향후 연구로는 코로나 확진자 추세에 영향을 끼칠 것으로 예상되는 독립 변수 추가와 연구의 한계점을 극복할 수 있는 예측 모델을 사용하는 것이다. 고려하고 있는 추가 독립 변수는 사회적 거리두기 운동의 단계와 시행 여부로, 8월 23일에 전국적으로 확대된 사회적 거리두기 운동이 시작된 이후로 확진자 수가 줄어드는 것을 <Figure 1>에서 확인할 수 있다. 본 연구가 갖는 한계점은 복잡한 딥러닝을 학습시키기에 데이터 개수가 적다는 점과 예측에 대한 신뢰도를 파악할 수 없다는 점이다. 따라서, 데이터 개수가 적은 상황에도 효과적으로 학습할 수 있고, 예측에 대한 신뢰도를 제공할 수 있는 베이지안 딥러닝 모델 등을 사용해 향후 연구를 이어가고자 한다.

## 참고문헌

Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V., and Alhyari, S. (2020). COVID-19 Prediction and Detection Using Deep Learning, *International Journal of Computer, Information Systems and Industrial Management Applications*, **12**, 168-181.

Arora, P., Kumar, H., and Panigrahi, B. K. (2020), Prediction and Analysis of COVID-19 Positive Cases Using Deep Learning Models

: A Descriptive Case Study of India, *Chaos, Solitons and Fractals*, **139**, 110017.
Chimmula, V. K. R. and Zhang, L. (2020), Time Series Forecasting of COVID-19 Transmission in Canada Using LSTM Networks, *Chaos, Solitons and Fractals*, **135**, 109864.
Friedman, J. H. (2002), Stochastic Gradient Boosting, *Computational Statistics and Data Analysis*, **38**(4), 367-378.
Grenfell, R. and Drew, T. (2020), Here's Why It's Taking So Long to Develop a Vaccine for the New Coronavirus, *Science Alert. Archived from the Original On*, 28.
He, S., Peng, Y., and Sun, K. (2020), SEIR Modeling of the COVID-19 and its Dynamics, *Nonlinear Dynamics*, **101**(3), 1667-1680.
Hochreiter, S. and Schmidhuber, J. (1997), Long Short-Term Memory, *Neural Computation*, **9**(8), 1735-1780.
Ministry of Health and Welfare (2020), COVID-19 cases, http://ncov.mohw.go.kr/tcmBoardList.do?brdId=3&brdGubun=.
Pavlov, Y. L. (2019), Random Forests, *Random Forests*, 1-122.
Rathore, S. S. and Kumar, S. (2016), A Decision Tree Regression based Approach for the Number of Software Faults Prediction, *ACM SIGSOFT Software Engineering Notes*, **41**(1), 1-6.
Zu, Z. Y., Di Jiang, M., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., and Zhang, L. J. (2020), Coronavirus Disease 2019(COVID-19) : A Perspective from China, *Radiology*, **296**(2), E15-E25.

## 저자소개

**배진수** : 건국대학교 수학과에서 2020년 학사학위를 취득하고 고려대학교 산업경영공학과에서 석・박통합과정에 재학중이다. 연구분야는 Realistic Semi-Supervised Learning, Bayesian Deep Learning이다.

**김성범** : 한양대학교 산업공학과에서 1999년 학사를 취득하고 2001년과 2005년 미국 Georgia Institute of Technology에서 산업공학 석사학위, 박사학위를 취득하였다. 미국 텍사스 주립대학교 교수를 역임하고 2009년부터 고려대학교 산업경영공학부 교수로 재직하고 있다. 연구분야는 인공지능, 머신러닝, 최적화이다.