Check for updates

Journal of the Korean Institute of Industrial Engineers
Vol. 49, No. 2, pp. 120-132, April 2023.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2023.49.2.120
© 2023 KIIE

코로나19 확진자 수 예측을 위한 딥러닝 모델: 확산 패러다임 변화를 반영한 입력변수 조정

홍태경[1] · 김은서[1] · 이희상[1,2†]

[1]성균관대학교 일반대학원 산업공학과 / [2]성균관대학교 시스템경영공학과

Deep Learning Models for Predicting Confirmed Cases of COVID-19: An Input Variable Conversion Technique Reflecting Virus Spread Patterns

Taekyung Hong[1] · Eunseo Kim[1] · Heesang Lee[1,2]

[1]Department of Industrial Engineering, Sungkyunkwan University
[2]Department of Systems Management Engineering, Sungkyunkwan University

It is difficult to predict with traditional approaches since the number of confirmed cases of COVID-19 fluctuates significantly. Therefore, this study tried to predict it by adopting machine learning models, such as Support Vector Regression (SVR), Random Forests (RF), eXtreme Gradient Boosting (XGBoost), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). Two experiments were performed with different prediction periods using two input data sets. The first input set consists of confirmed cases, severe cases, deaths, immunization cases, social distancing, and diffusion of COVID-19 mutations. The second input set is reconstructed to account for novel COVID-19 mutations without social distancing. As a result, the best deep learning model was selected for each experiment. This study showed that the model's input variables should be adjusted according to changes in the virus spread pattern. It also contributed to implementing policies to contain the COVID-19 spread.

Keywords: COVID-19, Multivariate Forecasting Model, Machine Learning, Deep Learning, Input Variable Conversion

1. 서 론

2020년 1월 30일 세계보건기구(WHO: World Health Organization)의 국제적 공중보건 비상사태(PHEIC) 발표 이래 신종코로나바이러스감염증-19(이하 코로나19)는 2022년 11월 시점에도 종식되지 못하고 있다. 일간 최대 코로나19 확진자(이하 확진자) 수는 미국과 EU에서는 100만 명 이상의 값을 기록하였던 반면 사우디아라비아, 남아프리카공화국에서는 4만을 넘지 않는 등 (Our World in Data, 2022), 국가마다 코로나19 확산의 규모 및 양상은 차이가 있었다. 이처럼 일간 확진자 수의 변화가 국가마다 다른 이유는 국가마다 바이러스 확산 요인의 양상과 그 강도가 변화하며 정확히 파악하기 어려운 데에 있다. 우리나라의 경우 2020년에는 지역집단발생으로 인한 지역사회 내 대규모 혹은 소규모 집단감염 사례가 지속되었고(Korea Disease Control and Prevention Agency, 2020), 2021년에는 지역사회접촉이 확진자의 주된 감염 경로로 확인되었다(Korea Disease Control and Prevention Agency, 2022). 또한 2021년 이후에는 바이러스의 변이가 코로나19의 전파력을 더욱 높이는 데에 일조하였으며,

이 논문은 2021년도 정부(과학기술보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2021R1F1A1063690).
† 연락저자 : 이희상 교수, 16419 경기도 수원시 장안구 서부로 2066, Tel : 031-290-7628, E-mail : leehee@skku.edu
2022년 10월 11일 접수; 2022년 11월 26일 수정본 접수; 2022년 11월 28일 게재 확정.

2022년 초에 우리나라를 포함한 여러 국가에서 전파력이 강한 오미크론 변이가 우세종으로 대두되며 바이러스의 확산이 가속화되었다. 다행히 국내 일간 확진자 수는 2022년 3월 17일을 기준으로 약 62만 명을 돌파한(Korea Disease Control and Prevention Agency, 2022) 것을 정점으로 2022년 11월 점차 누그러지고 있지만 아직도 변동성이 큰 실정이다. 본 연구는 딥러닝 모델을 활용하여 우리나라의 일간 확진자 수를 예측하는 것이다.

일간 확진자 수가 국가마다 다른 또 하나의 이유는 바이러스 방역을 위한 정부개입의 정도와 효과가 국가마다 다르기 때문으로 판단된다. 우리나라 정부는 스마트폰 앱과 빅데이터를 활용하여 바이러스 확산에 선제적으로 대응하였으나(Watson *et al.*, 2020; Kim, 2020), 확산세가 꺾이지 않자 사회적 거리두기를 시행하여 감염을 원천적으로 차단하고자 하였다. 사회적 거리두기는 거리두기 단계를 격상할 시 발생하는 사회적 혼란, 의료체계 실정에 부합하지 않는 격상 기준(Ministry of Health and Welfare, 2020), 감염 확산의 양상(Ministry of Health and Welfare, 2021) 등을 고려하여 개편을 거듭하며 다양한 형태로 시행되었다.

코로나19 백신 접종은 사회적 거리두기와 더불어 대표적인 방역 정책이다. 일반적으로 코로나19 백신은 항체 생성 및 유지를 위해 일정 시간 간격을 두고 두 번 이상 접종되는데, 우리나라의 1차 백신 접종은 2021년 2월 26일에 고령층 집단시설을 중심으로 시작되었다(Yoon, 2021). 2차 백신 접종이 2021년 7월부터 뒤늦게 본격적으로 이루어짐에 따라, 정부가 초기 백신 확보에 민첩하게 반응하지 못하였다는 비판이 일었으나(Kim, 2021) 2021년 9월 7일을 기준으로 전 국민의 59.9%, 35.8%에 해당하는 사람들이 백신 1차, 2차 접종을 완료하는 등(Ministry of Health and Welfare, 2021) 다른 나라에 비교하여 접종 시작은 늦었으나 다수의 국민에게 신속하게 이루어졌다. 백신 보급에 따라 항체를 가진 사람들의 비율이 늘면서 확진자 비율도 점차 감소하였으나, 재감염에 따라 백신 접종 수와 확진자 수는 유사하게 변화하였다.

또한 변이 바이러스의 출현도 확진자 수 변동에 영향을 주었다. 몇몇 변이는 전파력, 중증도 또는 백신, 치료제, 진단 도구 및 기타 공중보건, 사회적 조치 등의 효능에 영향을 미칠 수 있다(WHO, 2022). WHO는 이처럼 전파력 향상 및 유해한 역학적 변화가 확인되거나, 병원성 증가 또는 임상 질환 발현에 변화가 확인되거나, 진단, 백신, 치료제 등의 유효성 감소가 확인된 변이 바이러스들을 우려 변이 바이러스(Variants of Concern)로 지정하여 집중적으로 감시하였다(Korea Disease Control and Prevention Agency, 2022). WHO는 2020년 12월 18일에는 알파 변이와 베타 변이를, 2021년 1월 11일에는 감마 변이를, 2021년 5월 11일에는 델타 변이를, 2021년 11월 26일에는 오미크론 변이를 우려 변이 바이러스로 지정하였다(WHO, 2022).

코로나19 확산에 관한 데이터는 집계 및 제공 시점에 차이가 있는 경우가 빈번하다. 일례로 우리나라의 코로나19 검사 수 데이터는 코로나19 발원일 이래로 공공데이터포털을 통해 제공되었으나 2021년 11월부터 제공되지 않고 있다. 이와 같은 데이터 정책의 변화는 2021년 5월부터 편의점에서도 코로나19 자가진단키트를 구매할 수 있어(Lee, 2021) 국가 차원에서 개인의 검사 기록을 추적하기 어려웠고 단계적 일상회복이 시행되며 사망, 위중증, 병상가동률 등이 주요 지표로 선정되었기 때문이지만 확진자 수의 예측에는 어려움을 준다. 또한, 코로나19 확산 초기에는 나타나지 않거나 중요하지 않다고 판정되었던 데이터가 확산이 전개됨에 따라 중요해지기도 한다. 예를 들면 병상가동률은 2022년 2월 28일 이후의 값만 알 수 있으므로 그 이전의 예측에 사용하기 어려운 데이터가 된다.

확진자 수를 예측하는 본 논문에서 사용한 데이터는 대부분 전국 단위로 수집되는 일간 데이터이며, 당일 0시를 기준으로 집계된 데이터를 중심으로 사용하였다. 주 단위의 확진자 수 예측은 일 단위로 변화하는 확진자 수의 변동성을 망라하기 어려우므로 본 연구에서는 예측의 단위를 일간으로 결정하였다. 확진자 수 예측에 관한 선행 연구 중 구획 모델(compartment model), 시계열 모델(time series model) 등을 활용한 연구는 예측의 정확성이 낮고(Li *et al.*, 2021; Radha and Balamuralitharan, 2020; Satrio *et al.*, 2020; Shahid *et al.*, 2020; Cooper *et al.*, 2020), 다수의 선행 연구가 해외 각국의 확진자 수와 오미크론 변이의 확산 이전 시기의 예측에 집중하였다는(Luo *et al.*, 2021; Omran *et al.*, 2021; Rauf *et al.*, 2021; Wilson, 2021; Chen *et al.*, 2020; Chimmula and Zhang, 2020) 한계점을 고려하여 본 논문에서는 입수 가능한 데이터와 머신러닝 모델인 서포트 벡터 회귀(SVR: Support Vector Regression), 랜덤 포레스트(RF: Random Forests), 익스트림 그래디언트 부스팅(XGBoost: eXtreme Gradient Boosting)과 딥러닝 모델인 순환신경망(RNN: Recurrent Neural Network), 장기-단기 기억 신경망(LSTM: Long Short-Term Memory), 게이트 순환 유닛(GRU: Gated Recurrent Unit)을 활용하여 우리나라 실정에 부합하는 예측을 수행하였다. 본 논문에서 사용된 6가지 모델 모두 시계열 예측에 사용되는 모델로서 해외 확진자 수 예측에 활발하게 적용되고 있다. 본 논문의 예측 모델 개발 및 관련한 전산 실험은 코로나19의 확산 시기에 따라 확진자 추세와 더불어 확진자 수와 데이터의 상관관계가 달라지고 데이터 입수 가능 여부 자체도 달라진다는 점을 고려하여 설계하였으며, 오미크론 변이 발생 이전 기간과 전체 분석 기간에서 예측을 수행, 비교하여 각 기간에 대해 예측 정확도를 개선하고자 하였다. 코로나19 방지대책은 코로나19 확산에 따라 유동적으로 변화하는 양상을 보이므로, 일간 확진자 수의 정확한 예측은 사회적 거리두기 단계의 조정과 여타 방역 대책 정책의 실행에 일조하며 코로나19가 유발하는 불확실성에 대응할 수 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 확진자 수 예측에 관한 선행 연구를 분석하였고, 제3장에서는 논문에서 활용한 방법론과 데이터, 예측 수행 방법을 설명하였다. 제4장에서는 예측 결과를 확인, 비교하고 가장 우수한 예측을 제시하였으며 제5장에서는 본 논문이 학술적으로 기여하는 바와 더불

어 논문의 개선 방향을 제시하였다.

2. 선행 연구

확진자 수 예측에 사용되는 모델의 종류에는 크게 구획 모델, 시계열 모델, 머신러닝 및 딥러닝이 있다. 먼저 구획 모델을 사용한 확진자 수 예측 연구를 살펴보면 SIR(Susceptible-Infected-Recovered) 모델을 사용한 연구(Wilson, 2021; Ndiaye *et al.*, 2020), SEIR(Susceptible-Exposed-Infected-Recovered) 모델을 사용한 연구(Wu *et al.*, 2020)가 대표적이다. 구획 모델은 특정 지역의 전체 인구가 일정하다는 전제하에 전체 인구를 특성에 따라 나누어 예측을 수행하며, 이를 사용한 확진자 수 예측 연구들은 다른 모델들과의 성능 비교를 수행하지 않은 경우가 다수인 것이 확인되었다. 또한, 이론에 근거한 전통적 구획 모델의 특성상 변종이 돌출하는 코로나19 확산 실정에 적합하지 않은 "모델에 사용되는 일부 파라미터가 시간에 따라 불변한다(time-invariant)"라는 전제에 근거하고, 위중증 환자 수나 백신 접종 수, 감염병 확산 방지를 위한 정부의 대책 등 감염병 확산에 관한 주요 변수를 고려하기 어렵다는 한계가 있다(Ioannidis *et al.*, 2022; Moein *et al.*, 2021). 이를 개선하기 위해 기존 인구 구분을 세분화하는 구획 모델이 등장하였다(Crokidakis, 2020; Long *et al.*, 2020). 또한 시간적 요소를 추가한 모델(Li *et al.*, 2021; Chen *et al.*, 2020; Radha and Balamuralitharan, 2020), 전통 구획 모델의 가정을 변형한 모델(Cooper *et al.*, 2020; Siraj *et al.*, 2020) 등이 제시되며 구획 모델을 사용한 연구는 꾸준히 진행되고 있다.

시계열 모델은 시간에 따른 추세, 계절성을 제거하여 정상성을 갖는 시계열에 대해 예측을 수행하는 모델로 확진자 수 예측에 주로 사용되는 모델은 자기회귀누적이동평균(ARIMA: Auto-Regressive Integrated Moving Average) 모델 등 하나의 시계열 변수를 다루는 단변량 모델이다. 이 모델은 분석 대상인 시계열의 평균과 분산이 일정할 것을 가정하므로 평균 혹은 분산이 일정하지 않은 시계열은 연이은 관측값들의 차이인 차분(differencing) 혹은 시계열의 범위를 조정하는 변환(transformation)을 각각 적용한 후에 분석할 수 있다(Box *et al.*, 2016). 확진자 수 예측에 관한 다수의 연구에서는 확진자 수에 2차 이하의 차분을 적용하였으며(Wang *et al.*, 2021; Awan and Aslam, 2020; Sahai *et al.*, 2020; Kufel, 2020), 로그 변환을 적용하는 연구들도 발표되었다(Benvenuto *et al.*, 2020; Satrio *et al.*, 2020). 확진자 수 예측에 대해 시계열 모델을 사용한 예측이 상대적으로 우수하다는 연구도 존재하지만(Alabdulrazzau *et al.*, 2021) 시계열 모델은 예측값인 확진자 수의 증감에 영향을 미칠 수 있는 외부요인들을 예측에 충분히 반영하지 못하여 정확한 예측을 수행하기 어렵다는 한계를 가지며(Petropoulos *et al.*, 2022) 이는 구획 모델이 갖는 한계점과 유사하다.

머신러닝 및 딥러닝은 데이터의 패턴을 분석하여 정형 및 비정형 데이터 모두 처리할 수 있다. 그중 시계열 예측에 대표적인 모델은 RNN으로, 활성화 함수에 기반하는 딥러닝 모델로서 시계열이 갖는 비선형적 특성을 파악할 수 있고 구획 모델, 시계열 예측 모델보다 다양한 변수들을 사용하여 예측을 수행할 수 있다. 확진자 수의 예측은 RNN이 가진 그래디언트 소멸 문제를 보완한 LSTM을 사용하는 연구가 활발하다. LSTM의 전파 방향을 조정하거나 LSTM 셀을 누적하는 연구(Atik, 2022; Chandra *et al.*, 2022; Devaraj *et al.*, 2021)와 더불어 다른 머신러닝 및 딥러닝 모델과 결합하는 연구(Ayoobi *et al.*, 2021; Abbasimehr and Paki, 2021; Said *et al.*, 2021)가 활발하게 진행되었다. LSTM을 통해 캐나다의 코로나19 확산 고점 시기(Chimmula and Zhang, 2020), 중동 각국의 확진자 수(Alassafi *et al.*, 2022; Kafieh *et al.*, 2021) 등에 관한 예측이 실제 값에 근사한다는 연구 결과가 발표되기도 하였다. 이와 더불어 LSTM과 RNN, GRU 등 여타 RNN 계열 모델 혹은 SVR, RF, XGBoost 등의 모델과의 성능을 비교한 연구(Luo *et al.*, 2021; Omran *et al.*, 2021; Rauf *et al.*, 2021; Shahid *et al.*, 2020; Zeroual *et al.*, 2020)가 진행되었고, 대다수의 연구는 과거 확진자 수를 모델의 주요 학습 데이터로 이용하였다. 이러한 머신러닝 및 딥러닝을 사용한 예측은 확진자 수 예측에 가장 활발하게 적용되며, 구획 모델 혹은 시계열 모델을 사용할 때보다 더 정확한 것으로 확인되었다(Zoltar, 2021). 각 연구에서 제시하는 모델은 국가와 분석 기간에 따라 성능이 다른 것으로 확인되었다.

우리나라의 확진자 수 예측에 머신러닝을 최초로 사용한 연구는 5일 전까지의 과거 확진자 수와 법정 공휴일 여부에 최대최소변환을 적용하여 4일 후 확진자 수를 예측한 연구이다(Bae and Kim, 2021). 이 연구는 지역감염 및 집단감염 등의 확산 양상에 따라 7개의 실험 기간을 설정하고, LSTM, RF, XGBoost로 우리나라에서 발생하는 확진자 수의 전반적인 패턴의 변화를 예측한 결과, LSTM의 예측이 전반적으로 우수하였다. 이후의 연구로 양방향 LSTM(Bi-LSTM: Bidirectional LSTM) 및 GRU를 비교한 연구에 따르면 GRU의 성능이 더 우수하였으나, 2020년 1월부터 2021년 10월까지 기간에서 우리나라에서 발생하였던 확진자 수만을 다루어서 급증하거나 급감하는 확진자 수를 올바르게 예측하기 어려운 한계를 가지고 있다(Kim and Kim, 2022). 두 연구는 과거 확진자 수를 예측에 주로 사용하며 확진자 수에 영향을 미칠 수 있는 다른 사회적 요인들을 고려하지 않았던 반면, 예측 단위를 서울의 확진자 수로 설정하였던 연구는 확진자 수 외에도 사회적 거리두기 단계, 단어 "코로나"의 검색량, 지하철로 이동하는 인구수, 백신 접종 수, 날씨 데이터 등의 다양한 데이터를 사용하였다(Noh *et al.*, 2022). 이 논문에서는 LSTM으로 과거 5일간의 데이터로 서울시의 1일, 7일, 14일 후의 확진자 수를 예측하였는데, LSTM은 RNN, RF, XGBoost, ARIMA, 그래디언트 부스트 머신(GBM: Gradient Boost Machine), 라이트 GBM(LightGBM: Light Gradient Boost Machine), 다층 퍼셉트론(MLP: Multi-Layer Perceptron) 및 Bae and Kim의 연구에서 제시된 모델보다 우수하였다.

우리나라 확진자 수 예측에 관한 선행 연구들은 바이러스 확산 초반인 2020년 혹은 2021년에 집중하였으므로, 이들의 모델로는 2022년 3월에 급격히 증가하는 확진자를 파악하기 어렵다는 한계를 가졌다. 이에 본 논문에서는 머신러닝 및 딥러닝 모델을 이용하여 2020년부터 2022년에 이르기까지, 확진자 수와 관련된 다양한 공공데이터를 사용하되 코로나19 확산 양상에 따라 모델에 투입하는 입력변수를 바꾸고 하이퍼파라미터를 최적화하여 가장 적합한 모델을 선정하여 우리나라에서 발생하는 일일 확진자 수 예측을 시도하는 모델을 제시하고 실험 결과를 보고하였다.

3. 연구 방법

3.1 예측 모델

본 논문에서는 3가지 머신러닝 모델(SVR, RF, XGBoost)과 3가지 딥러닝 모델(RNN, LSTM, GRU)을 사용하여 일일 확진자 수를 예측하였다. 모든 모델은 python 3.8을 사용하여 구현하였으며, 모델의 우수성을 판단하기 위해 모든 실험은 5회 반복 시행한 후, 예측 평균값의 MAPE를 계산하여 비교하였다.

SVR은 데이터의 분류, 패턴 인식 및 다양한 종류의 데이터에 대한 회귀분석을 위해 자주 사용되는 머신러닝 알고리즘인 서포트 벡터 머신(Support Vector Machine)을 사용한 회귀분석 기법으로 각 데이터를 서로 다른 초평면(hyperplane)으로 분류하여 예측을 수행한다(Drucker *et al.*, 1996). SVR은 정규분포를 띤 입력 데이터를 전제하므로 예측 대상을 제외한 모든 입력변수는 정규화가 요구된다(Crone *et al.*, 2006). 이에, 본 논문에서는 SVR의 입력변수에 최대최소정규화(Min-Max Normalization)를 적용하였다. RF와 XGBoost는 여러 개의 학습기를 조합하여 사용하는 앙상블 학습(ensemble learning)으로, RF는 배깅(bagging), XGBoost는 그래디언트 부스팅(gradient boosting)을 접목한 기법이다(Breiman, 2001; Lee and Sun, 2020). 두 기법 모두 의사결정나무(decision tree)를 기저 학습기로 사용한다. 배깅은 부트스트랩(bootstrap) 표본 추출을 이용하여 복수의 데이터 집합을 구성한 후 각 학습기의 결과를 결합(aggregating)하는 기법이고, 그래디언트 부스팅은 부스팅 기법과 가중치를 반복적으로 갱신하는 경사 하강법(gradient descent)을 접목하는 기법이다(Chen and Guestrin, 2016).

선행 연구들을 참고하여 SVR, RF, XGBoost의 하이퍼파라미터는 <Table 1>과 같이 설정하였다.

RNN은 순차 데이터에 특화된 딥러닝 기법이며 시계열 예측, 기계번역, 감성 분류, 이미지 처리 및 음성 인식 등에 활발하게 이용되고 있다. 본 논문에서 차용한 RNN의 구조는 다대일(many-to-one)로, 일련의 과거 데이터를 토대로 일간 확진자 수를 예측한다. RNN은 입력층과 은닉층, 출력층으로 구성되며 인공신경망의 파라미터를 공유하고 시간에 따라 정보를 선별적으로 선택한다. RNN에서 t 시점(time step)에서의 은닉층은 t 시점의 입력 데이터와 $(t-1)$ 시점의 은닉 상태를 입력으로 받는다.

Table 1. Hyperparameters of Machine Learning Models

Model	Hyperparameter	Value
SVR	kernel	'rbf'
	gamma	'scale'
	C	1.0
	epsilon	0.1
RF	n_estimators	100
	criterion	'squared_error'
	min_samples_split	2
	min_samples_leaf	1
	max_features	n_features / 3
XGBoost	booster	'gbtree'
	eta(learning rate)	0.3
	max_depth	6
	min_child_weight	1
	sampling_method	'uniform'
	tree_method	'auto'
	base_score	0.5

RNN은 은닉층에서 과거 정보를 전달받아 가중치를 업데이트하며 학습을 진행하는데, 이 과정에서 그래디언트가 소멸할 수 있고 직전 시점의 정보만을 다음 예측에 사용하므로 부정확한 예측을 수행할 수 있다. 이를 해결하기 위해 LSTM과 GRU가 개발되었다. LSTM은 메모리 셀을 활용하여 RNN보다 그래디언트가 오래 지속될 수 있는 구조로, 메모리 셀은 입력 게이트(input gate, i_t)와 출력 게이트(output gate, o_t), 망각 게이트(forget gate, f_t)로 구성된다. 현재 시점인 t 시점에서, LSTM은 세 종류의 게이트를 이용하여 t 시점의 입력 데이터와 이전 시점인 $(t-1)$ 시점의 은닉 데이터, 셀 상태를 입력받아 t 시점의 은닉 데이터와 셀 상태를 출력한다. 입력 게이트는 t 시점의 메모리 셀이 t 시점의 입력 데이터와 $(t-1)$ 시점의 셀 상태를 얼마나 사용할지 결정하고(식 (1), (2)), 망각 게이트는 t 시점의 메모리 셀이 $(t-1)$ 시점의 셀 상태를 얼마나 사용할지를 결정한다(식 (3)). 이때 입력 게이트와 망각 게이트의 출력으로 t 시점의 셀 상태를 계산함으로써, 과거의 불필요한 정보를 제거한다(식 (4)). 출력 게이트는 t 시점의 은닉 데이터가 t 시점의 셀 상태를 얼마나 사용할지를 결정하며 t 시점의 은닉 데이터를 출력한다. 각 식에서 W_{xh_i}, W_{hh_i}, W_{xh_f}, W_{hh_f}, W_{xh_g}, W_{hh_g}, W_{xh_o}, W_{hh_o}는 파라미터이고, b_i, b_f, b_g, b_o는 편향이다.

$$i_t = \sigma(W_{xh_i}x_t + W_{hh_i}h_{t-1} + b_i) \quad (1)$$

$$g_t = ReLU(W_{xh_g}x_t + W_{hh_g}h_{t-1} + b_g) \quad (2)$$

$$f_t = \sigma(W_{xh_f} x_t W_{hh_f} h_{t-1} + b_f) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (4)$$

$$o_t = \sigma(W_{xh_o} x_t + W_{hh_o} h_{t-1} + b_o) \quad (5)$$

$$h_t = o_t \odot ReLU(c_t) \quad (6)$$

GRU는 리셋 게이트(reset gate, r_t)와 업데이트 게이트(update gate, z_t)로 구성되며, LSTM과 다르게 메모리 셀을 사용하지 않아 비교적 단순하다. t 시점의 리셋 게이트는 t 시점의 은닉 데이터 후보($\widetilde{h_t}$)가 $(t-1)$ 시점의 은닉 데이터를 얼마나 이용할지를 결정하고(식 (7)), 업데이트 게이트는 t 시점의 은닉 데이터 후보와 $(t-1)$ 시점의 은닉 데이터를 볼록 결합하여 각각을 t 시점의 은닉 데이터에 얼만큼을 반영할지를 결정한다(식 (8)). 각 식에서 W_r, U_r, W_z, U_z, W, U는 파라미터이고 b_r, b_z, b는 편향이다.

$$r_t = \sigma(W_r h_{t-1} + U_r x_t + b_r) \quad (7)$$

$$z_t = \sigma(W_z h_{t-1} + U_z x_t + b_z) \quad (8)$$

$$\widetilde{h_t} = ReLU(W(r_t \odot h_{t-1} + U x_t + b) \quad (9)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \quad (10)$$

딥러닝 모델의 경우 RNN, LSTM, 혹은 GRU를 은닉층으로 갖는 모델을 구현하였다. RNN, LSTM, GRU의 은닉 데이터의 차원 수(hidden size)는 2의 제곱수인 16, 32, 64 중 하나를 선택하되 많은 은닉층의 수를 사용할수록 낮은 차원을 선택하도록 하였으며 최대 3개의 은닉층을 고려하였다(<Table 2>). 또한, RNN과 더불어 LSTM과 GRU를 구성하는 게이트의 연산에는 기본 활성화함수와 순환 단계에서 사용하는 활성화함수가 요구되는데, Alassafi *et al.*(2022)의 연구 등을 참고하여 정류 선형 유닛(ReLU: Rectified Linear Unit)을 기본 활성화함수로, 시그모이드(sigmoid, σ) 함수를 순환 단계에서 사용하는 활성화함수로 설정하였다. ReLU를 활성화함수로 사용할 경우 출력값이 0으로 수렴할 수 있으므로 이를 방지하기 위해 균등 분포를 활용한 He 초기값을 커널의 초기값으로 설정하였다(He *et al.*, 2015). 모델의 손실 함수로 평균 제곱 오차(MSE: Mean Squared Error)를 설정하였으며 전체 학습의 반복 횟수와 배치 크기는 각각 500과 16으로 설정하였다. 파라미터 최적화의 경우, Zeroual *et al.*(2020), Devaraj *et al.*(2021) 등의 연구와 같이 아담 옵티마이저(adam optimizer)를 사용하였다.

Table 2. Hidden Size and Hidden Layers of RNN, LSTM, and GRU

Type of Layer	Hidden size	The Number of Hidden Layers
RNN/LSTM/GRU	64	1
	32	2
	16	3

3.2 데이터

예측을 위해 질병관리청 코로나바이러스감염증-19 공식 홈페이지(http://ncov.mohw.go.kr/), 한국언론진흥재단이 국내 주요 54개 언론사에서 보도하는 뉴스 데이터를 종합하여 제공하는 뉴스 빅데이터 시스템인 BigKinds에서 코로나19 확산에 관한 데이터를 수집하였다.

질병관리청 홈페이지에서 수집한 데이터는 일간 확진자의 발생 및 사망 현황, 일간 코로나19 예방접종 현황, 주간 코로나19 변이 검출현황, 사회적 거리두기 현황이다. 코로나19 변이 검출현황 데이터는 WHO에서 코로나19에 대한 우려 변이 바이러스로 지정하였던 알파, 베타, 감마, 델타, 오미크론 변이 각각의 주간 검출 건수이다. 일간 확진자 수 예측을 위해, 주간 데이터인 코로나19 변이 검출현황 데이터는 7로 나눈 후 각 주에 해당하는 일자들의 일간 데이터로 변환하였으며, 특정 요일의 여부는 0과 1로 부호화하였다.

사회적 거리두기는 지자체마다 그 강도와 시행 기간의 측면에서 차이가 있을 뿐만 아니라 서로 다른 두 지역의 사회적 거리두기 단계가 같더라도 시행 내용이 서로 다른 경우도 발생하였다. 따라서, 모든 지자체에서 공통으로 제한하였던 식당·카페 영업 가능 시간과 사적 모임 가능 인원을 기준으로 전국 각지에서 시행하였던 사회적 거리두기의 단계를 1부터 5까지의 수로 재구성하였다. 이를 위해 전국을 권역 단위(수도권, 충청권, 호남권, 경북권, 경남권, 강원, 제주)로 구분한 후, 권역별 사회적 거리두기 단계 및 강도, 시행 기간을 확인하였다. 같은 행정구역에 속하는 하부 지역들의 사회적 거리두기 강도는 지역별 확산 상황에 따라 다를 수 있으나 본 논문에서는 사회적 거리두기 단계를 하나의 전국적인 변수로 설정하였으므로, 하부 지역의 사회적 거리두기 기준은 무시하고 각 권역에 적용된 거리두기 기준을 고려하였다. 이때 사회적 거리두기가 하루 중에도 시간에 따라 변화하는 문제를 해결하기 위해 필요한 수정을 하였다. 예를 들면 2021년 7월 12일부터 2021년 9월 5일까지, 수도권의 사적모임 가능 인원수는 18시 이전에는 4명, 18시 이후에는 2명이었으므로 시간에 따라 비례배분하여 3.5명으로 계산하였다. 사회적 거리두기가 권역별로 다르지만 예측하고자 하는 확진자 수와 같이 전국적인 변수로 사용하기 위해서도 비슷한 조정이 필요하였다. 즉, 사회적 거리두기 단계를 권역별로 재구성한 후, 지역의 인구수를 기준으로 비례배분하여 모든 권역의 수치를 합산하여 전국의 1일 단위의 거리두기 단계를 재구성하였다. 재구성 기준은 <Table 3>과 같으며 재구성한 값은 제재의 강도와 비례한다.

BigKinds에서 수집한 데이터는 일간 코로나19 관련 뉴스의 수로, 뉴스 검색 시스템을 통해 단어 '코로나19', '코로나', '코로나 바이러스', '신종 코로나바이러스', 'COVID-19', '코비드19' 중 한 개 이상의 단어를 포함하는 뉴스의 수를 일별로 합산

Table 3. The New Standards of Social Distancing Level

Opening Hours of Restaurant and Cafe	Available Number of People in Private Meeting	Reconstructed Value of Social Distancing Level
no limit	no limit	1
until pm 9:00	49 or 99	2
until am 12:00	4 or 6 or 8 or 10 or 12	3
until pm 9:00 or pm 10:00	4 or 6 or 8	4
until pm 9:00 or pm 10:00	3.5 or 4	5

하여 확보하였다.

탐색적 데이터 분석(Exploratory Data Analysis)을 통해 각 데이터의 유형과 분포를 확인한 결과 특정 요일 여부 데이터는 이진형이었다. 확진자 수, 사망자 수, 입원환자 수, 관련 뉴스 검색 수는 정수형이었으며, 변이 검출 수, 사회적 거리두기 데이터는 실수형이었다. 일간 확진자 수는 2020년 2월 23일 이후부터 100을 상회하고, 결측치가 없었으므로 2020년 2월 23일을 예측의 시작 시점으로 분석 기간을 설정하였다. 이때 오미크론 변이 검출 확인 이후 확진자 수가 급증하였으므로 분석범위를 오미크론 변이 발생 이전 확산기(2020.02.23.~ 2021.11.27., 644일)와 전체 확산기(2020.02.23.~2022.04.30., 798일)로 구분하여 각 기간에 대해 다른 모델을 구현하였다. '실험 1'은 오미크론 변이 발생 이전 확산기에 관한 예측을, '실험 2'는 전체 확산기에 관한 예측을 수행하였다.

3.3 예측 수행

예측 수행을 위해 수집한 데이터는 변수로 처리하였다. 각 변수와 확진자 수와의 상관계수 절댓값이 0.6 이상이고 유의확률(p-value)이 0.05 이하로 유의한 변수를 예측에 사용하되, 코로나19 잠복기가 최대 14일이라는 점을 고려하여(Central Disease Control Headquarters & Central Disaster Management Headquarters, 2021) 특정 과거 시점으로부터 최대 14일 전까지의 코로나19 확진자 수 또한 변수로 고려하였다. 상관관계 확인 시 정수형 및 실수형 변수의 경우 피어슨 상관계수(pearson correlation coefficient)를 이진형 변수의 경우 점이연 상관계수(point-biserial correlation coefficient)를 사용하였다.

코로나19 변이 검출현황 데이터 등 실측치가 짧은 기간에 대해 존재하는 일부 데이터를 모델의 학습에 반영하기 위해 모든 모델의 학습 데이터와 테스트 데이터의 비율은 9:1로 설정하였으며 예측 기간이 다른 모델들을 비교하고자, 모델의 성능 평가의 측도로 데이터 단위의 영향이 작은 평균 절대 백분율 오차(MAPE: Mean Absolute Percentage Error)를 채택하였다(식 (11)). 식 (11)에서 n은 데이터 수, y는 실제값, $\hat{y}$는 예측값이다.

$$MAPE(\%) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \tag{11}$$

(1) 실험 1: 오미크론 변이 발생 이전 확산기(2020.02.23.~ 2021.11.27.)

"실험 1: 오미크론 변이 발생 이전 확산기"에서는 머신러닝 모델 SVR, RF, XGBoost와 딥러닝 모델 RNN, LSTM, GRU의 예측 성능을 비교하였으며, 실험 1의 모든 모델은 공통적으로 확진자 수, 사망자 수, 위중증 환자 수, 델타 변이 검출 수와 사회적 거리두기 변수를 예측에 사용하였다. 실험 1에서 고려한 모든 변수와 확진자 수 간의 상관계수 및 유의확률과, 이를 기준으로 선택한 변수를 정리하면 <Table 4>와 같다.

SVR과 XGBoost의 경우 예측에 반영하고자 하는 기간만큼

Table 4. Variable Selection in Experiment 1

Variable	Correlation Coefficient	p-value	Selected Variable in Experiment 1
Deaths	0.610	0.000	Selected
Severe Cases	0.862	0.000	Selected
Reconstructed Value of Social Distancing Level	0.600	0.000	Selected
The Number of News Articles	-0.358	0.000	Not Selected
Detected Alpha Variants	0.129	0.001	Not Selected
Detected Beta Variants	-0.022	0.569	Not Selected
Detected Gamma Variants	0.092	0.000	Not Selected
Detected Delta Variants	0.851	0.000	Selected
Pfizer Vaccination - 1st	0.485	0.000	Not Selected
Pfizer Vaccination - 2nd	0.524	0.000	Not Selected
Pfizer Vaccination - 3rd	0.505	0.000	Not Selected

Table 4. Variable Selection in Experiment 1(Continued)

Variable	Correlation Coefficient	p-value	Selected Variable in Experiment 1
Moderna Vaccination - 1st	0.384	0.000	Not Selected
Moderna Vaccination - 2nd	0.412	0.000	Not Selected
Moderna Vaccination - 3rd	0.409	0.000	Not Selected
AstraZeneca Vaccination - 1st	-0.019	0.640	Not Selected
AstraZeneca Vaccination - 2nd	0.262	0.000	Not Selected
AstraZeneca Vaccination - 3rd	0.593	0.000	Not Selected
Janssen Vaccination - 1st & 2nd	0.026	0.510	Not Selected
Janssen Vaccination - 3rd	0.459	0.000	Not Selected
Confirmed Cases 1 Day Ago	0.976	0.000	Selected
Confirmed Cases 2 Days Ago	0.952	0.000	Selected
Confirmed Cases 3 Days Ago	0.943	0.000	Selected
Confirmed Cases 4 Days Ago	0.937	0.000	Selected
Confirmed Cases 5 Days Ago	0.936	0.000	Selected
Confirmed Cases 6 Days Ago	0.949	0.000	Selected
Confirmed Cases 7 Days Ago	0.957	0.000	Selected
Confirmed Cases 8 Days Ago	0.935	0.000	Selected
Confirmed Cases 9 Days Ago	0.910	0.000	Selected
Confirmed Cases 10 Days Ago	0.898	0.000	Selected
Confirmed Cases 11 Days Ago	0.889	0.000	Selected
Confirmed Cases 12 Days Ago	0.890	0.000	Selected
Confirmed Cases 13 Days Ago	0.903	0.000	Selected
Confirmed Cases 14 Days Ago	0.909	0.000	Selected
If Today is Monday or Not	-0.065	0.100	Not Selected
If Today is Tuesday or Not	-0.063	0.110	Not Selected
If Today is Wednesday or Not	0.040	0.306	Not Selected
If Today is Thursday or Not	0.039	0.329	Not Selected
If Today is Friday or Not	0.030	0.442	Not Selected
If Today is Saturday or Not	0.035	0.371	Not Selected
If Today is Sunday or Not	-0.016	0.681	Not Selected

의 과거 확진자 수를 입력변수로 사용하였다. 데이터의 시계열적 특성을 고려하지 않는 RF의 경우 14일 전까지의 확진자 수를 입력변수로 사용하였다. 딥러닝 모델의 경우 time step을 고려하여 최대 14일 전 시점의 변수를 입력받되, 확진자 수의 경우 각 과거 데이터의 시점으로부터 time step일 전까지의 확진자 수를 사용하였다. 따라서 RF를 제외한 모든 모델은 예측에 반영하는 과거 시점의 수에 따라 14번 실험을 진행하며, 1일 전부터 최대 14일 전까지의 데이터로 당일 확진자 수를 예측하였다.

(2) 실험 2: 전체 확산기(2020.02.23.~2022.04.30.)

델타 변이와 오미크론 변이는 코로나19의 감염력이 극대화된 변이로, 우리나라에서도 두 변이가 등장하며 확진자 수가 급증하였다. 델타 변이는 2021년 7월 2주 이후로 변이 바이러스 분석 건수의 약 70%를 차지하며 우세종으로 대두되었으나, 오미크론 변이가 2022년 1월 3주 전체 변이 바이러스의 약 60%를 돌파함에 따라 검출 수가 점차 잦아들었다. 한편, 코로나19 팬데믹이 장기화됨에 따라 위드 코로나 정책이 널리 시행되며 사회적 거리두기의 강도는 점차 약화되었고, 정부 차원에서 백신 접종을 독려함으로써 10월 13일에는 18세 이상 성인을 기준으로 1차 접종률이 90.9%, 2차 접종 후 14일이 지난 백신접종완료자의 비율은 70.7%에 도달하여(Ministry of Health and Welfare, 2021) 백신 접종자 수가 크게 증가하는 등 확진자 수에 영향을 줄 수 있는 데이터에도 변화가 있었으므로 오미크론 변이 발생 이전 확산기에서 사용하였던 입력변수를 전체 확산기 예측에 사용할 경우 예측의 정확도가 낮아졌다. 따라서, 실험 2에서는 보다 정확한 예측을 위해 실험 1에서 사용하였던 입력변수 대신 실험 2 기간 중에 수집 가능한 데이터 중 확진자 수와의 상관계수 절댓값이 0.6 이상이고 유의확률이 0.05 이하로 유의한 데이터를 새로운 입력변수로 새롭게

Table 5. Variable Selection in Experiment 2

Variable	Correlation Coefficient	p-value	Selected Variable in Experiment 2
Deaths	0.893	0.000	Selected
Severe Cases	0.660	0.000	Selected
Reconstructed Value of Social Distancing Level	0.187	0.000	Selected
The Number of News Articles	-0.199	0.000	Not Selected
Detected Alpha Variants	-0.132	0.000	Not Selected
Detected Beta Variants	-0.108	0.002	Not Selected
Detected Gamma Variants	-0.078	0.027	Not Selected
Detected Delta Variants	-0.138	0.000	Not Selected
Detected Omicron Variants	0.724	0.000	Selected
Pfizer Vaccination - 1st	-0.101	0.004	Not Selected
Pfizer Vaccination - 2nd	-0.097	0.006	Not Selected
Pfizer Vaccination - 3rd	0.037	0.291	Not Selected
Pfizer Vaccination - 4th	0.150	0.000	Not Selected
Moderna Vaccination - 1st	-0.060	0.091	Not Selected
Moderna Vaccination - 2nd	-0.064	0.071	Not Selected
Moderna Vaccination - 3rd	0.008	0.828	Not Selected
Moderna Vaccination - 4th	0.151	0.000	Not Selected
AstraZeneca Vaccination - 1st	-0.059	0.097	Not Selected
AstraZeneca Vaccination - 2nd	-0.052	0.142	Not Selected
AstraZeneca Vaccination - 3rd	0.079	0.025	Not Selected
Janssen Vaccination – 1st & 2nd	-0.035	0.324	Not Selected
Janssen Vaccination - 3rd	-0.039	0.271	Not Selected
Janssen Vaccination - 4th	0.038	0.281	Not Selected
Novavax Vaccination - 1st	0.551	0.000	Not Selected
Novavax Vaccination - 2nd	0.666	0.000	Selected
Novavax Vaccination - 3rd	0.664	0.000	Selected
Novavax Vaccination - 4th	0.096	0.006	Not Selected
Confirmed Cases 1 Day Ago	0.957	0.000	Selected
Confirmed Cases 2 Days Ago	0.931	0.000	Selected
Confirmed Cases 3 Days Ago	0.928	0.000	Selected
Confirmed Cases 4 Days Ago	0.917	0.000	Selected
Confirmed Cases 5 Days Ago	0.918	0.000	Selected
Confirmed Cases 6 Days Ago	0.931	0.000	Selected
Confirmed Cases 7 Days Ago	0.938	0.000	Selected
Confirmed Cases 8 Days Ago	0.899	0.000	Selected
Confirmed Cases 9 Days Ago	0.864	0.000	Selected
Confirmed Cases 10 Days Ago	0.858	0.000	Selected
Confirmed Cases 11 Days Ago	0.839	0.000	Selected
Confirmed Cases 12 Days Ago	0.830	0.000	Selected
Confirmed Cases 13 Days Ago	0.839	0.000	Selected
Confirmed Cases 14 Days Ago	0.827	0.000	Selected
If Today is Monday or Not	-0.040	0.254	Not Selected
If Today is Tuesday or Not	0.004	0.914	Not Selected
If Today is Wednesday or Not	0.027	0.443	Not Selected
If Today is Thursday or Not	0.020	0.581	Not Selected
If Today is Friday or Not	0.003	0.930	Not Selected
If Today is Saturday or Not	0.003	0.934	Not Selected
If Today is Sunday or Not	-0.009	0.808	Not Selected

선정하였다. 그 결과, 실험 1에서 사용하였던 사회적 거리두기 변수와 델타 변이 검출 수는 전체 확산기에서의 확진자 수와 유의한 상관관계가 없으므로 입력변수에서 제외하였고, 확진자 수와 유의한 양의 상관관계가 있는 오미크론 변이 검출 수와 노바백스 백신 2차, 3차 접종자 수를 새로 고려하게 되었다. 즉, "실험 2: 전체 확산기"의 최종 입력변수는 코로나19 사망자 수, 위중증 환자 수, 오미크론 변이 검출 수, 노바백스 백신 2차 접종 수, 노바백스 백신 3차 접종 수, 그리고 과거 확진자 수이다. 실험 2에서 고려한 모든 변수와 확진자 수 간의 상관계수 및 유의확률과, 이를 기준으로 선택한 변수를 정리하면 <Table 5>와 같다.

과거 데이터를 모델 학습에 투입하여 예측을 수행하는 방식은 실험 1에서와 동일하며, 실험 2의 예측 결과를 실험 1에서 사용하였던 입력변수들을 실험 2에 사용하였을 때의 예측 결과와 비교하였다.

4. 연구 결과

4.1 실험 1: 오미크론 변이 발생 이전 확산기(2020.02.23.~2021.11.27.)

<Table 6>과 <Table 7>은 실험 1에서 예측에 반영하고자 하는 과거 기간(time period)에 따른 XGBoost, SVR 예측의 MAPE와, time step 및 은닉층 수별 RNN, LSTM, GRU 예측의 MAPE를 나타낸다. 시계열적 요소를 고려하지 않은 RF의 경우 예측의 MAPE는 23.21%로 계산되며 머신러닝 모델 중 가장 우수한 예측을 수행하는 모델은 RF로 나타났다.

머신러닝 모델과 딥러닝 모델 각각의 평균 MAPE는 42.98%, 15.29%로, 딥러닝 모델의 예측 평균이 비교적 정확하다는 것

Table 6. Experiment 1: Prediction Accuracy of Machine Learning Models (MAPE, %)

Time Period	XGBoost	SVR
1	29.74	78.89
2	30.96	78.80
3	30.72	78.63
4	24.75	78.69
5	25.39	78.60
6	25.80	78.55
7	25.57	78.63
8	26.27	78.73
9	25.77	78.78
10	26.26	78.94
11	26.32	79.00
12	25.83	79.13
13	26.27	79.21
14	26.70	79.29
Min	24.75	78.55
Max	30.96	79.29
Avg	26.88	78.85

Table 7. Experiment 1: Prediction Accuracy of Deep Learning Models (MAPE, %)

Time Step	1-layer RNN	2-layer RNN	3-layer RNN	1-layer LSTM	2-layer LSTM	3-layer LSTM	1-layer GRU	2-layer GRU	3-layer GRU
1	19.94	20.54	18.81	20.24	19.40	19.56	18.94	20.24	18.68
2	18.47	18.74	17.40	17.61	17.80	18.31	17.50	17.05	17.58
3	15.19	15.86	14.59	15.97	17.03	18.83	16.69	16.77	17.31
4	12.34	12.02	12.34	13.38	14.56	17.92	13.04	14.47	14.43
5	12.56	12.14	11.79	12.51	14.88	15.48	11.96	12.96	13.77
6	12.76	12.94	12.23	13.53	14.19	20.80	12.63	13.23	13.66
7	13.26	13.39	12.40	13.72	16.13	17.63	13.08	12.79	12.76
8	12.71	12.94	12.47	13.18	15.81	19.74	12.74	13.44	12.11
9	12.91	12.41	12.65	13.23	18.51	19.29	13.89	13.40	12.86
10	12.77	12.08	11.70	14.27	23.13	25.12	12.95	13.25	13.22
11	11.91	13.29	12.26	14.42	20.18	21.51	12.32	12.66	14.02
12	13.03	13.23	11.75	14.16	18.68	30.22	12.35	14.22	14.33
13	12.01	11.80	12.82	15.50	18.10	27.95	13.29	14.30	14.23
14	12.25	11.90	12.17	14.23	22.28	27.10	12.17	12.49	15.34
Min	11.91	11.80	11.70	12.51	14.19	15.48	11.96	12.49	12.11
Max	19.94	20.54	18.81	20.24	23.13	30.22	18.94	20.24	18.68
Avg	13.72	13.80	13.24	14.71	17.91	21.39	13.83	14.38	14.59

이 확인되었다. 딥러닝 모델을 살피면 RNN, LSTM, GRU를 은닉층으로 사용한 모델의 최소 MAPE는 각각 11.70%, 12.51%, 11.96%로, 정확한 예측값을 도출하는 time step은 모델의 종류에 따라 상이하였다. 모든 모델을 종합한 결과, 가장 우수한 예측은 time step이 10일 때, 3개의 RNN을 은닉층으로 사용한 딥러닝 모델로 MAPE는 11.70%를 기록하였다. <Figure 1>의 파란색 그래프는 실제 확진자 수를 나타내며, 주황색 그래프는 가장 우수한 머신러닝 모델이 예측한 확진자 수를, 빨간색 그래프는 가장 우수한 딥러닝 모델이 예측한 확진자 수를 나타낸다.

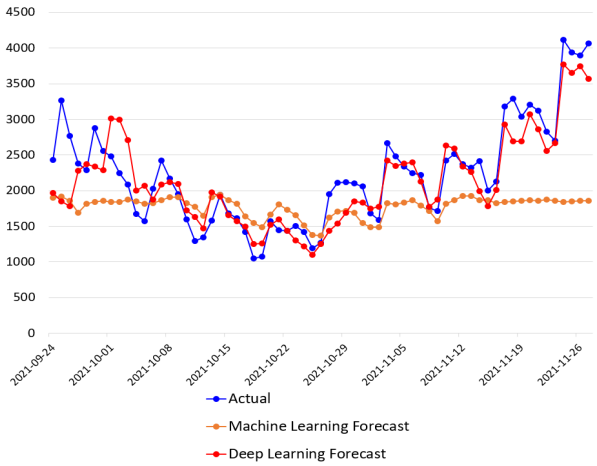

Figure 1. The Best Forecast in Experiment 1

4.2 실험 2: 전체 확산기(2020.02.23.~2022.04.30.)

실험 2에서는 실험 1에서 사용하였던 입력변수를 적용하되, 모든 모델의 하이퍼파라미터를 각각의 경우에 최적화한 후 결과를 비교하였다. 머신러닝 모델의 경우 time period를 최적화하였고, 딥러닝 모델의 경우 time step, 은닉층의 수, 은닉 데이터의 차원 수를 최적화하였다. 그 결과, 머신러닝 모델의 최소 MAPE는 각각 SVR 99.58%, XGBoost 62.49%, RF 65.62%로, 실험 1에서의 성능보다 크게 악화되었다. 따라서 실험 2에서는 머신러닝 모델의 사용은 더 이상 고려하지 않았다.

딥러닝 모델의 경우 실험 1에서 사용한 입력변수를 변화 없이 실험 2에 적용했을 때, 각 모델의 최소 MAPE는 RNN 24.88%, LSTM 28.25%, GRU 25.12%로 머신러닝 모델 예측의 MAPE보다 월등히 우수하였으나 실험 1에서 도출하였던 MAPE보다 컸다. 즉, 오미크론 변이 발생 이전 확산기인 실험 1에서 사용하였던 입력변수를 전체 확산기에도 계속해서 사용한다면 예측 정확도가 떨어지는 것을 알 수 있었다. 따라서 3.3.2절에서 설명한 것과 같이 새로운 입력변수를 사용하였다. <Table 8>은 실험 2에 새로운 입력변수를 사용하였을 때 3가지 딥러닝 모델의 MAPE를 나타내며, 각 모델의 MAPE는 time step과 모델에 따라 구분된다.

딥러닝 모델의 최소 MAPE는 RNN 18.44%, LSTM 24.04%, GRU 19.54%로, 실험 1에서 사용하였던 입력변수를 그대로 적용했을 때보다 우수한 성능을 보였다. 모든 모델을 종합한 결과, 실험 2에서 가장 우수한 예측은 time step이 14일 때, 1개의 RNN을 은닉층으로 사용한 딥러닝 모델로 MAPE는 18.44%를

Table 8. Experiment 2: Prediction Accuracy of Deep Learning Models (MAPE, %)

Time Step	1-layer RNN	2-layer RNN	3-layer RNN	1-layer LSTM	2-layer LSTM	3-layer LSTM	1-layer GRU	2-layer GRU	3-layer GRU
1	41.85	39.83	39.97	39.43	38.41	40.21	32.99	35.67	40.20
2	39.04	32.75	40.22	39.90	35.96	44.25	44.31	39.60	41.37
3	49.24	44.06	36.60	41.41	45.84	49.42	47.45	33.20	50.65
4	33.17	28.42	27.13	32.63	45.43	52.71	36.09	37.29	42.88
5	24.60	25.93	24.67	30.73	31.00	62.86	26.48	33.98	50.94
6	24.68	25.78	39.89	37.34	46.23	44.72	21.42	31.49	36.15
7	23.10	22.80	34.27	27.49	37.65	43.78	36.89	38.28	35.75
8	30.03	30.24	34.33	26.74	30.39	61.02	30.58	41.84	55.48
9	25.79	25.01	38.94	25.80	35.36	43.58	35.75	38.51	38.79
10	27.41	34.68	35.61	24.04	45.06	59.92	32.02	29.24	39.25
11	32.05	25.57	36.04	27.53	35.55	63.36	30.53	29.96	28.82
12	29.77	33.38	34.56	25.71	44.89	69.10	31.45	31.35	35.65
13	29.99	47.57	32.65	30.70	84.43	117.39	28.68	28.02	48.47
14	18.44	27.22	28.57	32.28	46.16	40.55	19.54	26.83	40.55
Min	18.44	22.80	24.67	24.04	30.39	40.21	19.54	26.83	28.82
Max	49.24	47.57	40.22	41.41	84.43	117.39	47.45	41.84	55.48
Avg	30.65	31.66	34.53	31.55	43.02	56.63	32.44	33.95	41.78

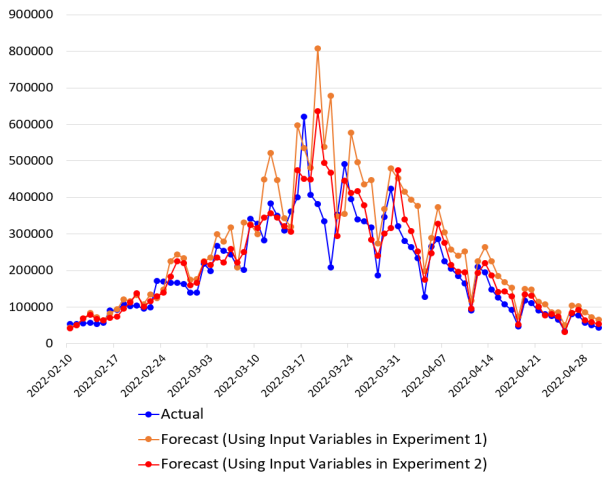

Figure 2. The Best Forecast Based on Input Variables in Experiment 2

기록하였다. <Figure 2>에서 파란색 그래프는 실제 확진자 수를 나타낸다. 주황색 그래프와 빨간색 그래프는 각각 실험 1에서의 입력변수를 사용하였을 때 가장 우수한 모델(time step이 6이고 3개의 RNN을 은닉층으로 사용하는 딥러닝 모델)이 예측한 확진자 수와, 실험 2에서 수정한 입력변수를 사용하였을 때 가장 우수한 모델(time step이 14이고 1개의 RNN을 은닉층으로 사용하는 딥러닝 모델)이 예측한 확진자 수를 나타낸다.

실험 2에서 RNN, LSTM, GRU로 확진자 수를 예측한 결과, 실험 1에서 사용하였던 입력변수를 사용할 때보다 외부 상황을 종합적으로 고려하여 입력변수를 전환하였을 때의 MAPE가 각각 6.44%p, 4.21%p, 5.58%p 개선되었다는 점을 알 수 있었다.

5. 결론 및 제언

본 논문은 머신러닝 및 딥러닝 기법과, 공공데이터를 사용하여 정확한 확진자 수를 예측하였다. 기존 연구들이 제시하였던 머신러닝 및 딥러닝 기반의 예측은 한정적인 데이터만을 사용하였으며 코로나19 확산 양상이 크게 변하지 않았던 초기에 한정하여 코로나19 확산 양상에 따라 변수의 중요도가 다를 수 있음을 반영하지 않아 예측이 부정확하다는 한계가 있었다. 본 논문은 지역과 시간에 따라 다른 사회적 거리두기 단계를 일관된 기준으로 재구성하여 사람들의 실제 생활에 직접적인 영향을 미치는 정책을 예측에 고려하였다. 또한 코로나19 변이 검출 수, 백신 접종 수 등 코로나19 확산과 관련하여 다양한 데이터를 단계별로 체계적으로 사용하였다. 특히 본 논문은 확진자 수 예측에서 중요하게 작용하는 입력변수와 최적 모델은 오미크론 변이 이전과 이후가 상이하다는 점을 주목하였다. 즉, 오미크론 변이 등장과 같이 코로나19 확산에 중대한 영향을 미치는 사건이 발생하였을 때, 분석 기간에 따라 입력변수 및 예측 모델을 다르게 적용하여 예측을 수행하면 예측의 정확도를 높일 수 있다는 점을 확인함으로써 우리나라의 코로나19 확산 실태와 예방 정책 실정에 부합하는 예측 수행에 기여하였다. 본 논문에서 제시하는 예측 모델이 코로나19 확산 패러다임의 전환에 따라 입력변수를 다르게 사용하여 성능을 제고한다는 장점은 다른 예측 문제에서도 모형의 단계적인 개발이 효과적일 수 있다는 시사점을 준다.

본 논문은 크게 두 가지 측면에서 개선될 수 있다. 첫째로 정책변수의 책정 기준을 보완할 수 있다. 현재 사용된 정책 변수는 전국 단위의 사회적 거리두기 변수밖에 사용하지 않았다. 향후 다중이용시설 유형별 방역수칙, 마스크 착용 의무화, 백신패스, 자가격리 의무화 등 다양한 정책을 예측에 반영할 수 있을 것이다. 둘째로 본 논문에서 이용하였던 데이터보다 양질의 데이터를 사용한다면 예측 성능을 지속적으로 개선할 수 있을 것이다. 본 논문에서 진행하였던 실험에는 기간 내 제공되던 도중 삭제 처리되는 등 데이터 완전성에 위배된 데이터는 사용하지 않았다. 또한 데이터 발표 기관 및 집계 시점의 차이로 인해 변동이 있는 데이터는 최신의 자료를 사용하여 데이터의 안정성을 고려하였으나 향후 연구에서는 보다 안정된 데이터의 확보가 바람직하다. 이동 및 의료 데이터와 같이 민감 데이터를 예측에 사용할 수도 있을 것이며, 국내 발생과 더불어 해외 유입과 관련된 데이터도 추가로 고려할 수 있을 것이다. 이처럼 사용할 입력 데이터를 충분히 확보한다면, 2022년 4월 이후로 감소 추세를 나타내다 7월 이후 다시 증가하는 등 향후에도 예측이 필요한 확진자 수의 변동 폭을 잘 포착하는 향상된 모델을 구축할 수 있을 것이다. 또한 지역별로 다른 데이터를 확보할 수 있다면 본 연구와 같은 전국 단위의 확진자 수 예측을 지역별 확진자 수 예측으로 확장하여, 지역 단위의 세밀한 예측을 수행할 수 있고 지자체의 실정에 맞는 확산 방지 대책에도 일조할 수 있을 것이다.

참고문헌

Abbasimehr, H. and Paki, R. (2021), Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization, *Chaos Solitons Fractals*, **142**, 110511.

Alabdulrazzaq, H., Alenezi, M. N., Rawajfih, Y., Alghannam, B. A., Al-Hassan, A. A., and Al-Anzi, F. S. (2021), On the accuracy of ARIMA based prediction of COVID-19 spread, *Results in Physics*, **27**, 104509.

Alassafi, M. O., Jarrah, M., and Alotaibi, R. (2022), Time series predicting of COVID-19 based on deep learning, *Neurocomputing*, **468**, 335-344.

Atik, I. (2022), COVID-19 Case Forecast with Deep Learning Bi-LSTM Approach: The Turkey Case, *International Journal of Mechanical Engineering*, **7**(1), 6307-6314.

Awan, T. M. and Aslam, F. (2020), Prediction of daily COVID-19 cases in European countries using automatic ARIMA model, *Journal of*

Public Health Research, **9**(1765), 227-233.

Ayoobi, N., Sharifrazi, D., Alizadehasni, R., Shoeibi, A., Gorriz, J. M., Moosaei, H., Khosravi, A., Nahavandi, S., Chofreh, A. G., Goni, F. A., Klemeš, J. J., and Mosavi, A. (2021), Time Series Forecasting of New Cases and New Deaths Rate for COVID-19 using Deep Learning Methods, *Results in Physics*, **27**, 104495.

Bae, J.-S. and Kim, S-B. (2021), Predictions of COVID-19 in Korea Using Machine Learning Models, *Journal of the Korean Institute of Industrial Engineers*, **47**(3), 272-279.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., and Ciccozzi, M. (2020), Application of the ARIMA model on the COVID-2019 epidemic dataset, *Data Brief*, **29**, 105340.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2016), *Time Series Analysis: Forecasting and Control*, Wiley.

Breiman, L. (2001), Random Forests, *Machine Learning*, **45**, 5-32.

Central Disease Control Headquarters and Central Disaster Management Headquarters (2021), *Guidelines for Responding to COVID-19 (for local governments)*, **10**, 1-288.

Chandra, R., Jain, A., and Singh Chauhan, D. (2022), Deep learning via LSTM models for COVID-19 infection forecasting in India, *PLoS One*, **17**(1), 1-28.

Chen, T. and Guestrin, C. (2016), XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Chen, Y. C., Lu. P. E., Chang, C. S., and Liu, T. H. (2020), A time-dependent SIR model for COVID-19 with Undetectable Infected Persons, *IEEE Transactions on Network Science and Engineering*, **7**(4), 3279-3294.

Chimmula, V. K. R. and Zhang, L. (2020), Time series forecasting of COVID-19 transmission in Canada using LSTM networks, *Chaos Solitons Fractals*, **135**, 109864.

Cooper, I., Mondal, A., and Antonopoulos, C. G. (2020), A SIR model assumption for the spread of COVID-19 in different communities, *Chaos Solitons Fractals*, **139**, 110057.

Crokidakis, N. (2020), Data analysis and modeling of the evolution of COVID-19 in Brazil, arXiv:2003.12150.

Crone, S. F., Lessmann, S., and Stahlbock, R. (2006), The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, *European Journal of Operational Research*, **173**(3), 781-800.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1996), Support Vector Regression Machines, *Advances in Neural Information Processing Systems*, **9**, 155-161.

Devaraj, J., Madurai Elavarasan, R., Pugazhendhi, R., Shafiullah, G. M., Ganesan, S., Jeysree, A. K., Khan, I. A., and Hossain, E. (2021), Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?, *Results in Physics*, **21**, 103817.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proceedings of the IEEE International Conference on Computer Vision*, **2015**, 1026-1034.

Ioannidis, J. P. A., Cripps, S., and Tanner, M. A. (2022), Forecasting for COVID-19 has failed, *International Journal of Forecast*, **38**(2), 423-438.

Kafieh, R., Arian, R., Saeedizadeh, N., Amini, Z., Serej, N. D., Minaee, S., Yadav. S. K., Vaezi, A., Rezaei, N., and Haghjooy Javanmard, S. (2021), COVID-19 in Iran: Forecasting Pandemic Using Deep Learning, *Computational and Mathematical Methods in Medicine*, **2021**, 6927985.

Kim, N. (2021, August), COVID-19: How did Korea become 'last place in vaccination' in 'K quarantine'?, BBC News Korea, Retrieved September 30, 2022, from https://www.bbc.com/korean/news-58313506.

Korea Disease Control and Prevention Agency (2020), Coronavirus Disease-19 (COVID-19) one-year outbreak major cluster infection report as of January 19, 2021, in the Republic of Korea, *Public Health Weekly Report*, **14**(9), 482-495.

Korea Disease Control and Prevention Agency (2021), Vaccination is carried out according to the vaccination plan in August~September, 1-41.

Korea Disease Control and Prevention Agency (2022), COVID-19 Variants, Retrieved September 29, 2022, from https://kdca.go.kr/contents.es?mid=a20107020000.

Korea Disease Control and Prevention Agency (2022), Current Status of COVID-19 Outbreak and Vaccination in Korea (2022.3.18.), 1-7.

Korea Disease Control and Prevention Agency (2022), Current Status of COVID-19 Outbreak in Korea (2022.6.28.), 1-18.

Korea Disease Control and Prevention Agency (2022), One-Year Report of COVID-19 Outbreak in the Republic of Korea, January-December 2021, *Public Health Weekly Report*, **15**(4), 225-234.

Korea Government (2021), Signed contract with Pfizer for additional 40 million doses of COVID-19 vaccine, 1-5.

Kufel, T. (2020), ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries, *Equilibrium*, **15**(2), 181-204.

Kim, J. H. and Kim, J. Y. (2022), Comparative analysis of performance of BI-LSTM and GRU algorithm for predicting the number of Covid-19 confirmed cases, *Journal of the Korea Institute of Information and Communication Engineering*, **26**(2), 187-192.

Kim, M. S. (2020, March), South Korea is watching quarantined citizens with a smartphone app, *MIT Technology Review*, Retrieved July 1, 2022, from https://www.technologyreview.com/2020/03/06/905459/coronavirus-south-korea-smartphone-app-quarantine/.

Lee, B. (2021, May), Corona self-test kit sold at convenience stores… Check the result in 30 minutes, Retrieved October 2, 2022, The JoongAng, from https://www.korea.kr/news/policyNewsView.do?newsId=148906366.

Lee, Y. -J. and Sun, J. -W. (2020), Predicting Highway Concrete Pavement Damage using XGBoost, *Korean Journal of Construction Engineering and Management*, **21**(6), 46-55.

Li, Y., Ge, L., Zhou, Y., Cao, X., and Zheng, J. (2021), Toward the Impact of Non-pharmaceutical Interventions and Vaccination on the COVID-19 Pandemic With Time-Dependent SEIR Model, *Frontier of Artificial Intelligence*, **4**, 648579.

Long, Y. S., Zhai, Z. M., Han, L. L., Kang, J., Li, Y. L., Lin, Z. H., Zeng, L., Wu, D. Y., Hao, C. Q., Tang, M., Liu, Z., and Lai, Y. C. (2020), Quantitative assessment of the role of undocumented infection in the 2019 novel coronavirus (COVID-19) pandemic, arXiv:2003.12028.

Luo, J., Zhang, Z., Fu, Y., and Rao, F. (2021), Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms, *Results in Physics*, **27**, 104462.

Ministry of Health and Welfare (2020), COVID-19 Central Disaster and Safety Countermeasure Headquarters Regular Briefing (2020.11.1.), 1-73.

Ministry of Health and Welfare (2021), Based on adults 18 years of age and older, the first dose of the COVID-19 vaccine is 90.9%, and the completion rate is 70.7%, 1-53.

Ministry of Health and Welfare (2021), Implementation of New Fourth Stage of Social Distancing in Metropolitan Area (7.12~ 7.25), 1-27.
Ministry of Health and Welfare (2021), The largest one-day vaccinations were carried out (1.36 million doses) yesterday, 1-57.
Moein, S., Nickaeen, N., Roointan, A., Borhani, N., Heidary, Z., Javanmard, S. H., Ghaisari, J., and Gheisari, Y. (2021), Inefficiency of SIR models in forecasting COVID-19 epidemic: A case study of Isfahan, *Scientific Reports*, **11**(1), 1-9.
Ndiaye, B. M., Tendengm L., and Seck, D. (2020), Analysis of the COVID-19 pandemic by SIR model and machine learning technics for forecasting, arXiv:2004.01574.
Noh, Y.-A., Jung, S.-W., Moon, J.-U., and Hwang, E. J. (2022), LSTM-based Daily COVID-19 Forecasting Scheme Considering Social Variables, *The Korean Institute of Information Scientists and Engineers*, **28**(2), 116-121.
Omran, N. F., Ghany, S. F. A., Saleh, H., Ali, A. A., Gumaei, A., and Al-Rakhami, M. (2021), Applying Deep Learning Methods on Time-Series Data for Forecasting COVID-19 in Egypt, Kuwait, and Saudi Arabia, *Complexity*, **2021**, 1-13.
Our World in Data (2022), Daily new confirmed COVID-19 cases, Retrieved June 28, 2022, from https://ourworldindata.org/explorers/coronavirus-data-explorer?facet=none&Metric=Confirmed+cases&Interval=New+per+day&Relative+to+Population=false&Color+by+test+positivity=false&country=USA~ITA~DEU~GBR~FRA~JPN~KOR~HKG~CAN~ZAF~RUS~MEX~BRA~SAU~ARG~European+Union~IND~IDN~CHN~TUR.
Petropoulos, F., Makridakis, S., and Stylianou, N. (2022), COVID-19: Forecasting confirmed cases and deaths with a simple time series model, *International Journal of Forecast*, **38**(2), 439-452.
Radha, M. and Balamuralitharan, S. (2020), A study on COVID-19 transmission dynamics: stability analysis of SEIR model with Hopf bifurcation for effect of time delay, *Advances in Different Equations*, **523**, 1-20.
Rauf, H. T., Lali, M. I. U., Khan, M. A., Kadry, S., Alolaiyan, H., Razaq, A., and Irfan, R. (2021), Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks, *Personal and Ubiquitous Computing*, 1-18.
Sahai, A. K., Rath, N., Sood, V., and Singh, M. P. (2020), ARIMA modelling & forecasting of COVID-19 in top five affected countries, *Diabetes & Metabolic Syndrome*, **14**(5), 1419-1427.
Said, A. B., Erradi, A., Aly, H. A., and Mohamed, A. (2021), Predicting COVID-19 cases using bidirectional LSTM on multivariate time series, *Environmental Science and Pollution Research*, **28**(40), 56043-56052.
Satrio, C. B. A., Darmawan, W. Nadia, B. U., and Hanafiah, N. (2020), Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET, *Procedia Computer Science*, **179**, 524-532.
Shahid, F., Zameer, A., and Muneeb, M. (2020), Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM, *Chaos Solitons Fractals*, **140**, 110212.
Wang, G., Wu, T., Wei, W., Jiang, J., An, S., Liang, B., Ye, L., and Liang, H. (2021), Comparison of ARIMA, ES, GRNN and ARIMA-GRNN hybrid models to forecast the second wave of COVID-19 in India and the United States, *Epidemiology and Infection*, **149**, 1-9.
Watson, I., Jeong, S., Hollingsworth, J., and Booth, T. (2020, March), How this South Korean company created coronavirus test kits in three weeks, CNN, Retrieved July 1, 2022, from https://edition.cnn.com/2020/03/12/asia/coronavirus-south-korea-testing-intl-hnk/index.html.
WHO (2022, September), Tracking SARS-CoV-2 variants, Retrieved September 29, 2022, from https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.
Wilson, D. J. (2021), Weather, Social Distancing, and the Spread of COVID-19. Federal Reserve Bank of San Francisco, *Federal Reserve Bank of San Francisco*, 1-36.
Wu, J. T., Leung, K., and Leung, G. M. (2020), Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study, *The Lancet*, **395**(10225), 689-697.
Yoon, S. (2021, February), February 26, Public vaccinations against COVID-19 launched nationwide, Korea.net, Retrieved September 30, 2022, from https://www.korea.net/NewsFocus/policies/view?articleId=195357&searchKey=all&searchValue=vaccination&pageIndex=1.
Zeroual, A., Harrou, F., Dairi, A., and Sun, Y. (2020), Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study, *Chaos Solitons Fractals*, **140**, 110121.
Zoltar (2021, November), COVID-19 US Forecast Evaluation, Retrieved November 12, https://covid19forecasthub.org/eval-reports/?state=US&week=2021-11-12.

저자소개

홍태경 : 성균관대학교 행정학과, 시스템경영공학과에서 2021년 학사학위를 취득하고 성균관대학교 산업공학과에서 2023년 석사학위를 취득하였다. 연구분야는 경영과학, 데이터마이닝이다.

김은서 : 성균관대학교 시스템경영공학과에서 2021년 학사학위를 취득하고 성균관대학교에서 산업공학과 석사과정에 재학 중이다. 연구분야는 데이터마이닝, 최적화이다.

이희상 : 서울대학교 산업공학과에서 학사학위와 석사학위를 취득하고 Georgia Tech에서 Industrial & Systems Engineering 박사학위를 취득하였다. KT 선임연구원, 한국외국어대학교 조교수/부교수를 역임하고 2004년부터 성균관대학교 시스템경영공학과에서 교수로 재직 중이다. 연구분야는 경영과학, 비즈니스 애널리틱스, 기술경영이다.