碩士學位 請求論文
指導敎授 李 熙 祥

딥러닝 모델을 활용한 국내 코로나19 확진자 수 예측

- 코로나19 확산에 따른 입력변수 전환 -

成均館大學校 一般大學院

産業工學科

洪 兌 炅

碩士學位請求論文

딥러닝 모델을 활용한 국내 코로나19 확진자 수 예측

2023

洪兌炅

碩士學位 請求論文
指導敎授 李 熙 祥

딥러닝 모델을 활용한 국내 코로나19 확진자 수 예측

- 코로나19 확산에 따른 입력변수 전환 -

Predicting Confirmed Cases of COVID-19 Using Deep Learning Models

- An Input Variable Conversion Technique Reflecting Virus Spread Patterns -

成均館大學校 一般大學院

産業工學科

洪 兌 炅

碩士學位 請求論文
指導敎授 李 熙 祥

딥러닝 모델을 활용한 국내 코로나19 확진자 수 예측

- 코로나19 확산에 따른 입력변수 전환 -

Predicting Confirmed Cases of COVID-19 Using Deep Learning Models

- An Input Variable Conversion Technique Reflecting Virus Spread Patterns -

이 論文을 工學 碩士學位請求論文으로 提出합니다.

2022 年 10 月 日

成均館大學校 一般大學院

産業工學科

洪 兌 炅

이 論文을 洪 兌 炅의 工學

碩士學位 論文으로 認定함.

2022 年 12 月 日

審査委員長

審査委員

審査委員

목차

표목차

그림목차

논문요약

딥러닝 모델을 활용한 국내 코로나19 확진자 수 예측

국내 신종코로나바이러스감염증-19(이하 코로나19) 확진자 수는 급격한 변동을 거듭하였다. 또한, 코로나19에 대한 데이터의 가용 여부가 시간에 따라 변화하였으므로 구획 모델 등의 전통적인 접근으로는 코로나19 확진자 수를 예측하기 어렵다. 이를 정확하게 예측하기 위해 본 논문에서는 다양한 과거 데이터를 입력변수로 사용하는 딥러닝 방법론을 채택하였으며, 분석 기간을 나눈 후 각 기간에서 가용한 데이터를 활용하였다. 본 논문에서 제시하는 딥러닝 예측 모델의 은닉층은 RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory), 또는 GRU(Gated Recurrent Unit)로 구성되며, 모델의 예측은 MAPE(Mean Absolute Percentage Error)로 평가하였다. 분석 기간은 코로나19 변이 검출을 기준으로 분할하여, 변이 발생 전 코로나19 확산기, 코로나19 변이 확산기, 오미크론 변이 확산기로 구분하며 기간마다 별도의 실험을 수행하였다.

모든 실험은 연령대별 확진자 수, 수도권·비수도권 확진자 수, 코로나19로 인한 사망자 수 및 재원 위중증 환자 수, 사회적 거리두기 단계, 과거 확진자 수를 입력변수로 고려하였다. 시간과 지역에 따라 강도가 다른 사회적 거리두기는, 바이러스 확산 방지를 위해 실시한 두 가지 조치를 기준으로 재구성하였다. 해외유입과 코로나19 확산 관련 뉴스 수의 경우, 각 실험에서 코로나19 확진자 수와의 상관관계가 유의할 경우 입력변수로 사용하였고 코로나19 변이 검출 수, 병상가동률, 재택치료자 수는 코로나19 변이 확산기 실험 혹은 오미크론 변이 확산기 실험에서 입력변수로 고려하였다. 각 실험에서 입력변수는 확진자 수와의 상관계수 크기에 따라 구분

하여, 어떤 변수를 사용할 때 예측 성능이 가장 우수한지 확인하였다.

분석 결과, 모든 실험에 걸쳐 가장 우수한 모델은 하나의 모델로 특정할 수는 없었으나 각 실험에서 가장 우수한 예측의 MAPE는 15% 이내로 계산되었다. 코로나19 변이 확산기에서 사용한 입력변수를 오미크론 변이 확산기에 적용하면 오미크론 변이 확산기에서 가용한 데이터로 입력변수를 구성할 때보다 예측 성능이 악화됨을 확인하였다. 본 연구는 바이러스 확산 패턴의 변화에 따라 모델의 입력변수를 조정해야 함을 보여주었으며, 일관된 데이터를 안정적으로 수급하기 위한 데이터 정책 정비의 필요성을 제시하였다.

주제어 : 코로나19 확진자, 다변량 시계열 예측, RNN, LSTM, GRU

제1장 서 론

1. 연구 배경

1) 신종코로나바이러스감염증-19 확산 추세

국내 일간 신종코로나바이러스감염증-19(이하 코로나19) 확진자 수는 2022년 3월 17일을 기준으로 약 62만 명을 돌파한(Korea Disease Control and Prevention Agency, 2022) 것을 정점으로 2022년 11월 점차 누그러지고 있다. 그림 1은 우리나라에서 발생한 일간 코로나19 확진자 수(이하 확진자 수)를 나타낸다. 미국과 EU에서 일간 최대 확진자 수는 100만 명 이상의 값을 기록하였으나 일정한 시계열 패턴은 발견되지 않았던 반면, 남아프리카공화국에서의 일간 최대 확진자 수는 4만에 이르지 못하지만 약 6개월을 주기로 하여 증감 추세가 변화하는(Our World in Data, 2022) 등, 코로나19 확산의 규모 및 양상은 국가마다 차이가 있었다. 그림 2는 인구 백만 명당 확진자 수를 시각화하여 각 대륙의 인구별 확진자 수를 나타낸 것으로, 인구수와 관계없이 대륙마다 확진 추세가 다양하며 모든 대륙에서 2021년 12월 이후 확진자 수가 급증하는 것을 알 수 있다. 일간 확진자 수의 변화가 급격하고 예측하기 어려운 이유는 바이러스 확산 요인의 양상을 정확히 파악하기 어려운 데에 있다. 일례로 우리나라의 경우 시간에 따라 바이러스 확산의 주된 요인이 변화하였다. 우리나라의 경우 2020년에는 지역집단발생으로 인한 지역사회 내 대규모 혹은 소규모 집단감염 사례의 지속이(Korea Disease Control and Prevention Agency, 2020), 2021년에는 지역사회접촉이 확진자의 주된 감염 경로로 확인되었다(Korea Disease Control and Prevention Agency, 2022).

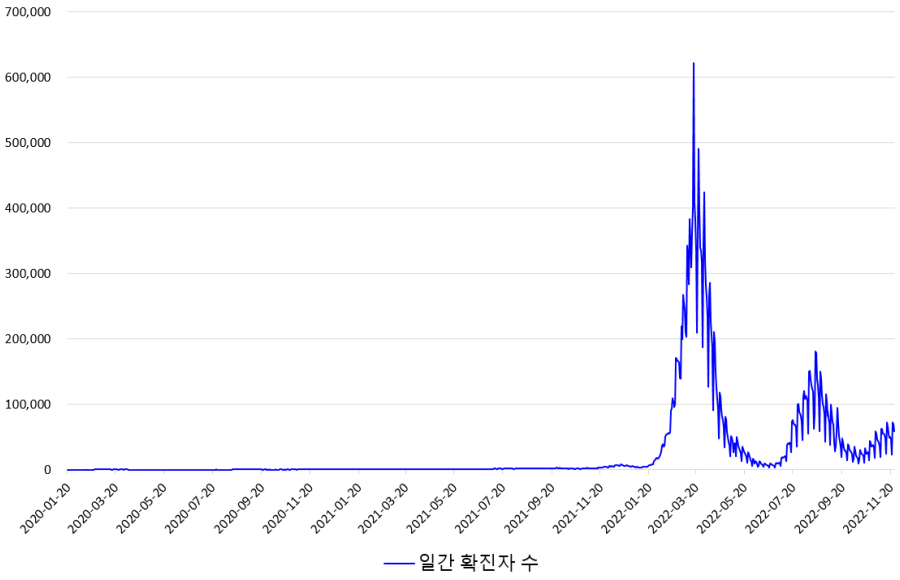

그림 1. 우리나라의 일간 확진자 수

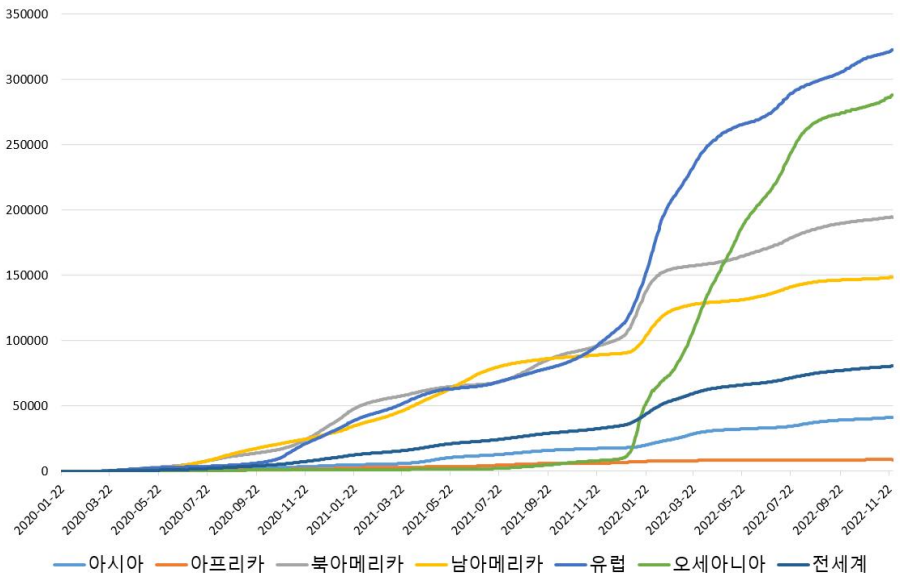

그림 2. 인구 백만 명당 일간 확진자 수

2) 코로나19 확산 방지를 위한 정부개입

일간 확진자 수가 국가마다 상이한 이유로는 바이러스 방역을 위한 정부개입의 정도와 그 효과가 정부마다 다르기 때문으로 판단되며, 가장 대표적인 정부개입에는 사회적 거리두기와 코로나19 백신 접종이 있다. 우리나라의 경우 스마트폰 앱과 빅데이터를 활용하여 바이러스 확산에 선제적으로 대응하였으나(Watson et al., 2020; Kim, 2020), 확산세가 꺾이지 않자 사회적 거리두기를 시행하여 사람들 사이의 대면 접촉을 최소화함으로써 감염을 원천적으로 차단하고자 하였다. 사회적 거리두기는 2020년 2월 29일에 공식 시행이 공표된 이래로(Center for Disease Control, 2020) 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 토대로 개편을 거듭하며 다양한 형태로 시행되었다.

일반적으로 코로나19 백신은 항체 생성 및 유지를 위해 일정 시간 간격을 두고 두 번 이상 접종된다. 정부는 '범정부 백신도입 TF'를 출범시켜 백신을 안정적으로 수급하고자 노력하였으나(Korea Government, 2021) 수급에 차질을 빚으며(Korea Disease Control and Prevention Agency, 2021) 본격적인 2차 접종은 2021년 7월부터 뒤늦게 이루어졌다. 이에 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) 시간이 지남에 따라 백신 접종은 신속하게 이루어졌다. 백신 보급에 따라 항체를 가진 사람들의 비율이 늘면서 확진자 비율이 점차 감소하였던 반면, 재감염이 확산되며 백신 접종 수와 확진자 수는 유사하게 증가하였다.

3) 코로나19 변이 바이러스

코로나19는 비말을 통해 감염되는 등 전파력이 강할 뿐만 아니라, 변이 속도가 매우 빨라 관련 백신 및 치료제 개발을 더디게 하였다. 이때 일부 변이는 전파력, 중증도 또는 백신, 치료제, 진단 도구 및 기타 공중보건, 사회적 조치 등의 효능에 영향을 미칠 수 있음이 확인되는 등(WHO, 2022) 확진자 수 변동에 영향을 줄 수 있다. 세계보건기구(WHO: World Health Organization)는 기존 바이러스보다 전파력 혹은 병원성이 높거나 해로운 역학적 변화가 확인된 변이 바이러스, 임상 질환 발현에 변화가 확인된 바이러스, 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). 2022년 11월까지 우려 변이 바이러스로 지정된 변이에는 알파 변이, 베타 변이, 감마 변이를, 델타 변이, 오미크론 변이가 있다(WHO, 2022).

2. 논문 개요 및 구성

코로나19 확산에 관한 데이터는 집계 및 제공 시점에 차이가 있는 경우가 빈번하다. 일례로 우리나라의 코로나19 검사 수 데이터는 코로나19 발원일 이래로 공공데이터포털(https://www.data.go.kr/)을 통해 제공되었으나 2021년 11월에 그동안의 데이터를 삭제 처리한 후 제공되지 않고 있다. 이는 2021년 5월부터 편의점에서도 코로나19 자가진단 키트를 구매할 수 있어(Lee, 2021) 국가 차원에서 개인의 자가진단 기록을 추적하기 어려웠고, 단계적 일상회복이 시행되며 사망, 위중증, 병상가동률 등이 주요 지표로 선정되었음에 기인한다. 이러한 데이터는 확진자 수 예측과 밀접한 관련이 있음에도 불구하고 안정적으로 확보할 수 없으므로 예측 모델에 투입하기가 어렵다. 또한, 코로나19 확산 초기에는 나타나지 않거나 중요하지 않다고 판정되었던 데이터가 확산이 전개됨에 따라 중요해지기도 한다. 병상가동률은 2022년 2월 28일 이후의 값만 알 수 있으므로 그 이전의 예측에 사용하기에는 어려운 데이터이었으나 2022년 2월 28일 이후에는 유용한 데이터가 된다.

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 & Balamuralitharan, 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 & Zhang, 2020) 한계점을 고려하여 본 논문에서는 입수 가능한 데

이터와 RNN, LSTM, GRU 등의 딥러닝 모델을 활용하여 우리나라 실정에 부합하는 예측을 수행하였다. 3가지 모델 모두 시계열 예측에 사용되는 모델로서 해외 확진자 예측에 활발하게 적용되고 있다. 실험은 시간에 따라 확진자 추세와 더불어 확진자와 유의한 상관관계가 있는 데이터가 달라진다는 점을 고려하여 설계하였다. 이러한 예 중 하나로 코로나19 방지대책은 코로나19 확산에 따라 유동적으로 변화하는 양상을 보이므로, 일간 확진자 수의 정확한 예측은 사회적 거리두기 단계의 조정과 여타 방역 대책 정책에 따라 달라진다는 사실이다. 따라서 본 연구는 코로나19 변이 발생을 기준으로 분석 기간을 세 개로 구분하여 예측을 수행, 비교하여 각 기간에 대해 정확한 예측을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 확진자 수 예측에 관한 선행 연구들을 분석하였고, 3장에서는 논문에서 활용한 방법론과 데이터, 예측 수행 방법을 설명하였다. 4장에서는 예측 결과를 확인, 비교하고 가장 우수한 예측을 제시하였으며 5장에서는 본 논문이 학술적으로 기여하는 바와 더불어 논문의 개선 방향을 제시하였다.

제2장 문헌 연구

1. 모델 분석

1) 구획 모델

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 바이러스에 감염될 위험이 있는 그룹(Susceptible), 바이러스에 감염된 그룹(Infected), 바이러스에서 회복되거나 바이러스로 인해 사망한 그룹(Recovered), 바이러스에 감염되었으나 전파력이 없는 그룹(Exposed) 등 특성에 따라 구분하여 예측을 수행한다. 구획 모델로 확진자 수를 예측한 선행 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, 재원 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 세분화하여 전체 인구를 면밀하게 구별하는 구획 모델이 등장하였다(Crokidakis, 2020; Long et al., 2020). 또한 시간적 요소를 추가한 모델(Li et

al., 2021; Chen et al., 2020; Radha & 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.

2) 시계열 모델

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 & 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 갖는다(Petropoulos et al., 2022). 이는 구획 모델이 갖는 한계점과 유사하다.

3) 머신러닝 및 딥러닝 모델

머신러닝 및 딥러닝 모델은 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 & Paki, 2021; Said et al., 2021)가 활발하게 진행되었다. LSTM을 통해 캐나다의 코로나19 확산 고점 시기(Chimmula & 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.

2. 머신러닝을 활용한 우리나라의 확진자 수 예측

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 & 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 우수하였다(Kim & Kim, 2022). 하지만 이 연구는 2020년 1월부터 2021년 10월까지 기간에서 우리나라에서 발생하였던 확진자 수만을 다루어서 이후에 급증하거나 급감하는 확진자 수를 올바르게 예측하기 어려운 한계를 가지고 있다(Kim & Kim, 2022). 두 연구가 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, 코로나19 백신 접종 수, 날씨 데이터 등의 다양한 데이터를 사용한 연구가 있다(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 & Kim의 연구에서 제시된 모델보다 우수하였다.

3. 모델 선정

Covid19 ForecastHub(https://covid19forecasthub.org/)는 전 세계 팀들이 미국 CDC(Centers for Disease Control and Prevention)와 협력하여 미국 전역 및 주 단위의 확진자 수 및 코로나19로 인한 사망자 수, 입원환자 수를 예측한 결과를 모은 웹사이트이다. 이 웹사이트에 게시된 예측 모델은 크게 구획 모델, 시계열 예측 모델, 통계적 모델, AI 예측 모델로 구분되며, 2021년 11월 12일에 제출된 모델은 구획 모델 7개, 시계열 예측 모델 4개, 통계적 모델 4개, AI 예측 모델 9개이다(COVID-19 ForecastHub, 2021). COVID-19 ForecastHub의 기준 모델인 'COVIDhub-baseline'과, 위 24개 모델의 상대적 평균 절대 오차(Relative MAE: Relative Mean Absolute Error) 모델에 대해 예측을 비교한 결과, 가장 우수한 모델은 AI 예측 모델이었다. 즉, 기준 모델과 비교 모델의 상대적 MAE가 1보다 작으면 비교 모델이 더 우수함을 뜻하는데(Morel et al., 2022), AI 예측 모델의 상대적 MAE는 0.83으로 네 유형의 모델 중 가장 작은 것으로 나타났으므로, 본 논문에서는 딥러닝 모델을 활용하여 예측을 수행하였다. 이때, 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. 따라서, 본 논문에서는 2020년부터 2022년에 이르기까지 확진자 수와 관련된 다양한 공공데이터를 사용하되, 코로나19 확산 양상에 따라 분석 기간을 나누어 모델에 투입하는 입력변수를 전환하였다. 특정 시점을 예측할 때 예측에 반영할 과거 데이터의 시점 수를 고정하였던 선행 연구와는 달리, 본 논문에서는 이를 하이퍼파라미터로 다루어 1일 전부터 최대 14일 전까지의 데이터를 고려함으로써 예측 성능을 높이고자 하였다. 또한, 본 논문은 모든 하이퍼파라미터를 최적화하여 가장 적합한 모델을 선정하여 우리나라에서 발생하는 일간 확진자 수를 예측하는 모

델을 제시하고 실험 결과를 보고하였다.

제3장 연구방법론

1. 예측 모델 수립

1) 예측 모델 개요

본 논문은 우리나라의 일간 확진자 수 예측을 위해 은닉층으로 RNN, LSTM, GRU를 활용하는 딥러닝 모델을 제안하며, 최대 14일 전까지의 과거 데이터를 토대로 다음 날 발생할 확진자 수 예측을 수행한다.

2) 순방향 인공신경망

순방향 인공신경망(Feed-forward Neural Network)은 딥러닝 모델의 기틀이 되는 모델로, 입력층과 은닉층, 출력층으로 구성된다. 각 층은 다음 층과 연결되며, 하나의 층에는 다수의 노드(node)가 존재한다. 순방향 인공신경망에 투입된 데이터는 입력층과 은닉층, 출력층을 차례로 거치며 역행하지 않는다. 먼저 입력층은 입력받은 데이터를 은닉층으로 전달하는 역할로, 입력층 내부를 구성하는 노드의 수는 입력받는 데이터의 개수와 동일하게 결정된다. 은닉층은 여러 개가 존재할 수 있는 유일한 층이며, 입력층 혹은 이전 출력층으로부터 전달받은 데이터에 활성화 함수를 적용하여 다음 은닉층 혹은 출력층으로 데이터를 전달한다. 활성화 함수에는 선형 활성화 함수와 비선형 활성화 함수가 있지만, 선형모델이 갖는 한계를 극복하기 위해 일반적으로 시그모이드(Sigmoid, σ), 하이퍼볼릭 탄젠트(Hyperbolic Tangent, tanh), 정류 선형 유닛(ReLU: Rectified Linear Unit, $ReLU$) 등의 비선형함수가 사

용된다. Sigmoid는 0 또는 1에 근사하므로 이와 같은 이진 문제에 자주 적용되며, tanh는 연산이 빠르고 -1 또는 1에 근사하여 sigmoid보다 활용도가 높다. 두 함수 모두 미분 시 0이 되는 지점이 존재하므로 그래디언트가 소멸할 위험이 존재한다. 이러한 문제를 해결하기 위해 ReLU가 고안되었으나, 입력값이 0보다 작은 경우 그래디언트가 소멸할 수 있다. ReLU는 0보다 큰 입력값을 그대로 출력한다는 점에서 sigmoid와 tanh보다 연산이 간편하다. 식 (1), (2), (3)은 각각 sigmoid, tanh, ReLU를 나타낸다.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2}$$

$$ReLU(x) = \max(0, x) \tag{3}$$

출력층은 이전 층에서 전달받은 데이터를 FNN이 해결해야 하는 문제의 목적에 맞게 출력하는 층으로, 마지막 은닉층이 전달한 값을 그대로 출력하기도 하지만 활성화 함수로 가공한 데이터를 출력하기도 한다. 일례로 입력받은 이미지 속 물체가 강아지인지 아닌지 구별하는 문제는 0 또는 1만 출력하도록 출력층을 조정하여 해결할 수 있다. 즉, 특정 이미지를 순방향 인공신경망에 투입하여 1이 출력된다면 해당 사진은 강아지 사진으로 판별할 수 있을 것이다.

3) RNN

일정한 순서가 있는 순차 데이터(sequential data)를 입력받는 RNN은 순방향 인공신경망의 은닉층에 재귀 기능을 더한 인공신경망으로 시계열 예측, 기계번역, 감

성 분류, 이미지 처리 및 음성 인식 등에 활발하게 이용되는 딥러닝 모델이다. 시계열 데이터를 입력받는 RNN에서, 입력층은 일정 기간 혹은 단일 시점(time step)의 과거 데이터를 입력받는다. t 시점의 은닉층은 t 시점의 입력 데이터와 t 시점의 바로 이전 과거 시점에 해당하는 $(t-1)$ 시점의 은닉 상태를 입력으로 받아 t 시점에서의 은닉 상태를 도출한다. 이를 수식과 도식으로 표현하면 식 (4)와 같다. h_t는 t 시점에서의 은닉 상태, W_h와 W_x는 은닉 상태와 입력 데이터의 파라미터로써 은닉층의 파라미터에 해당하고 b는 편향이다.

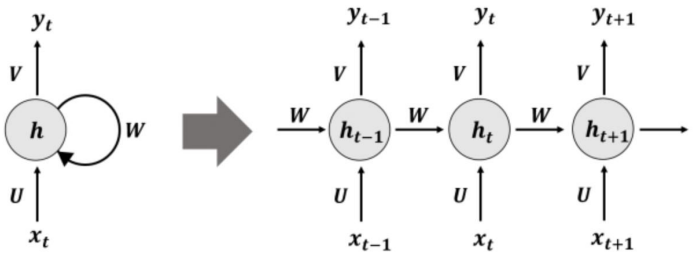

그림 3. RNN의 구조

$$h_t = \tanh(h_{t-1}W_h + x_tW_x + b) \tag{4}$$

RNN과 같은 인공신경망이 활용하는 데이터는 크게 학습 데이터와 테스트 데이터로 나뉜다. 학습 데이터를 인공신경망에 투입하여 인공신경망 내부의 파라미터를 조정 테스트 데이터를 통해 모델의 성능을 측정할 수 있다. RNN의 학습은 순방향 인공신경망의 학습과 동일하게, 출력값과 실제값의 오차에 관한 손실 함수를 최소화하는 방향으로 이루어지며 은닉층의 파라미터는 그래디언트를 활용한 역전파 과정을 통해 업데이트된다. 출력값과 실제값의 차이를 최소화하기 위해, 전파 과정은 출력층에서 은닉층, 입력층 방향으로 진행되고, 매 층을 지날 때마다 그래디언트를 계산하며 딥러닝 모델이 도출한 값과 실제 값 간의 차이를 각 노드에 전달한다. 은

닉층의 파라미터는 모든 시점의 은닉층과 공유할 뿐만 아니라, 은닉 상태는 과거 은닉 상태를 토대로 업데이트되므로 RNN은 과거 정보를 선택적으로 활용하여 예측을 수행한다는 특징이 있다. RNN은 입력받는 데이터와 출력하는 데이터의 시간 길이에 따라 다대일(many-to-one) RNN, 다대다(many-to-many) RNN, 일대일(one-to-one) RNN, 일대다(one-to-many) RNN 등으로 구분된다. 그 중, 다대일 구조는 여러 시점의 과거 데이터를 입력받아 단일 시점에서의 예측값을 출력하고, 다대다 구조는 여러 시점의 데이터를 입력받아 여러 시점에서의 예측값을 출력하는데, 입력받는 데이터의 시간 길이와 출력값의 시간 길이는 다를 수 있다. 일련의 과거 데이터를 토대로 일간 확진자 수를 예측하는 본 논문에서는 many-to-many RNN 및 many-to-one RNN을 차용하여 모델을 구성하였다. RNN이 은닉층으로서 기능할 경우 다대다 RNN을 사용하였고, RNN이 최종 예측값을 출력할 경우 다대일 RNN을 사용하였다. 각 모델을 도식화하면 다음과 같다.

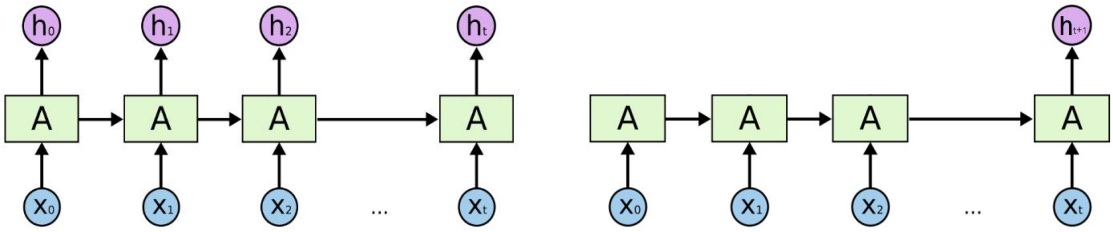

그림 4. Many-to-many RNN(좌), many-to-one RNN(우)

4) LSTM

RNN이 다루는 데이터의 시간 순서가 길수록, 역전파 과정에서 계산되는 그래디언트가 점차 감소하여 초기 입력값에 도달하기 전에 0으로 소멸할 위험이 있다. 그래디언트가 소멸할 경우 가중치는 업데이트되지 못하며, 최종적으로 모델의 성능이

악화된다. 이를 해결하기 위해 LSTM과 GRU가 개발되었다. LSTM은 메모리 셀(memory cell)과 셀 상태(cell state)를 활용하여 RNN보다 그래디언트가 오래 지속될 수 있는 딥러닝 모델이다. LSTM의 메모리 셀은 RNN의 노드와 대응하는 개념으로, 입력 게이트(input gate, i_t)와 망각 게이트(forget gate, f_t), 출력 게이트(output gate, o_t)로 구성된다. LSTM의 구조는 다음과 같다.

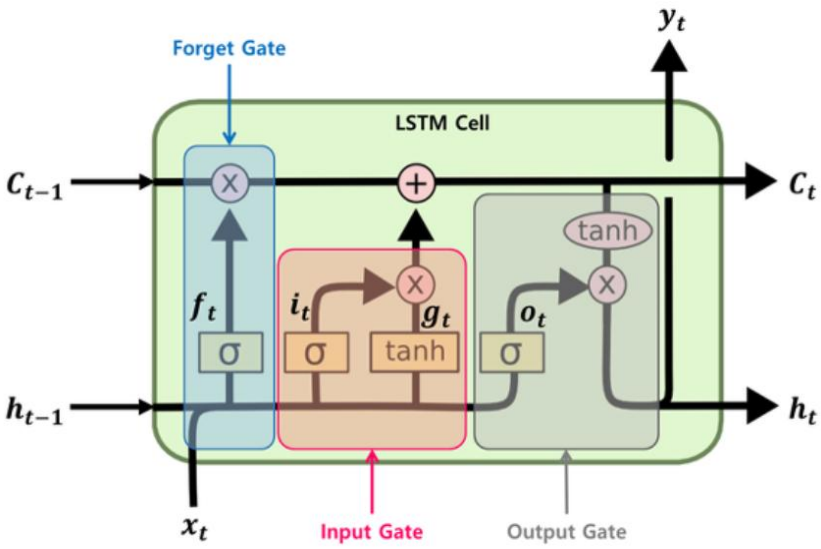

그림 5. LSTM의 구조

t 시점의 메모리 셀은 t 시점의 입력 데이터, $(t-1)$ 시점의 은닉 상태와 셀 상태를 입력받아 t 시점의 은닉 상태와 셀 상태를 출력한다. 각 게이트에서의 연산은 다음과 같다.

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

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

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

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

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

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

$$(0 \leq i_t, f_t, o_t \leq 1)$$

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. 먼저 입력 게이트는 t 시점의 메모리 셀이 t 시점의 입력 데이터와 $(t-1)$ 시점의 은닉 상태를 얼마나 사용할지 결정하고(식 (1), (2)), 망각 게이트는 t 시점의 메모리 셀이 $(t-1)$ 시점의 셀 상태를 얼마나 사용할지를 결정한다(식 (3)). 두 게이트를 통해 과거의 불필요한 정보를 제거하여 t 시점의 셀 상태를 계산한 후(식 (4)), 이를 얼마나 사용할지를 출력 게이트로 결정하여(식 (5)), t 시점의 은닉 상태를 출력한다(식 (6)). 식 (6)의 ⊙는 두 행렬의 성분을 각각 곱하는 연산인 아다마르 곱(Hadamard product)을 뜻한다.

5) GRU

GRU는 후보 은닉 상태($\widetilde{h_t}$)를 활용하고, 셀 상태를 사용하지 않아 LSTM보다 연산 속도가 빠르다. GRU는 리셋 게이트(reset gate, r_t)와 업데이트 게이트(update gate, z_t)로 구성되며 다음의 구조를 갖는다.

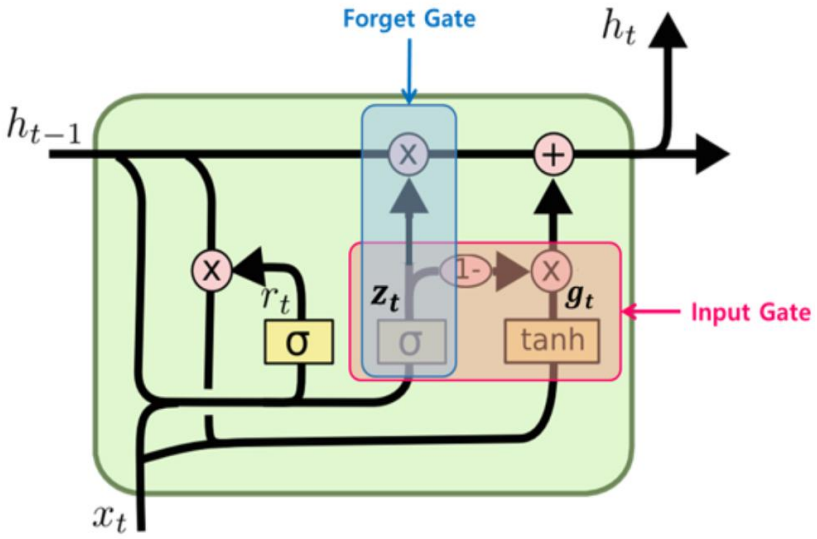

그림 6. GRU의 구조

t 시점의 GRU는 t 시점의 입력 데이터와 $(t-1)$ 시점의 후보 은닉 상태를 입력받아 t 시점의 후보 은닉 상태 및 은닉 상태를 출력한다. 각 게이트에서의 연산은 다음과 같다.

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

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

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

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

$$(0 \le r_t, z_t \le 1)$$

각 식에서 W_r, U_r, W_z, U_z, W, U는 파라미터이고 b_r, b_z, b는 편향이다. t 시점의 리셋 게이트는 t 시점의 후보 은닉 상태가 $(t-1)$ 시점의 은닉 상태를 얼마나 이용할지를

결정하고(식 (7)), t 시점의 업데이트 게이트는 t 시점의 후보 은닉 상태와 $(t-1)$ 시점의 은닉 상태를 볼록 결합하여, 각각을 t 시점의 은닉 상태에 얼마나 반영할지를 결정한다(식 (8)). 후보 은닉 상태 t 는 $(t-1)$ 시점의 은닉 상태를 바탕으로 t 시점에서 얻은 정보를 뜻하며(식 (9)), t 시점의 은닉 상태는 $(t-1)$ 시점의 은닉 상태와 t 시점의 후보 은닉 상태를 활용하여 계산된다(식 (10)).

6) 하이퍼파라미터 설정

본 논문에서는 RNN, LSTM, 혹은 GRU를 은닉층으로 갖는 모델을 구현하여, 모든 모델의 하이퍼파라미터인 time step, 은닉층의 수, 은닉 상태의 차원 수를 최적화한 후 각 결과의 우수성을 비교하였다. RNN, LSTM, GRU의 은닉 상태의 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. RNN과 LSTM, GRU의 활성화 함수에는 일반적으로 tanh를 사용하지만 Alassafi et al.(2022)의 연구 등을 참고하여 ReLU를 사용하였으며, ReLU의 출력값이 0으로 수렴하는 것을 방지하기 위해 균등 분포를 활용한 He 초기값을 가중치의 초기값으로 설정하였다(He et al., 2015). 이는 입력층의 노드 수를 고려하여 가중치의 범위를 결정한다(식 (11)). 모델의 손실 함수로 평균 제곱 오차(MSE: Mean Squared Error)를 설정하였으며(식 (12)), 전체 학습의 반복 횟수와 배치 크기는 각각 500과 16으로 설정하였다. 파라미터 최적화의 경우, Zeroual et al. (2020), Devaraj et al. (2021) 등의 연구에서 활용된 바와 같이 수렴 성능과 속도가 우수한 아담 옵티마이저(Adam Optimizer)를 사용하였다.

$$W \sim U(-\sqrt{\frac{6}{n_{in}}}, \sqrt{\frac{6}{n_{in}}}) \quad (11)$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2 \quad (12)$$

식 (11)에서 n_{in}은 입력층의 노드 수를 뜻하고, 식 (12)에서 n은 데이터의 개수, $\hat{y_i}$는 모델의 출력값, y_i는 예측하고자 하는 값을 뜻한다.

7) 상관분석

본 논문에서는 딥러닝 모델에 투입할 입력변수를 선별하기 위해 상관분석을 실시하였다. 상관분석은 상관계수(r)를 계산하여 서로 다른 두 변수(X, Y) 사이의 연관관계를 분석하는 방법론으로, 대략적인 연관관계는 두 변수에 대한 산점도를 그려 파악할 수 있다. 상관계수는 -1부터 1까지의 값을 가지며, 양수이면 두 변수의 관계가 양의 상관관계임을 뜻한다. 이는 한 변수가 증가할 때 다른 변수는 감소함을 나타낸다. 반면 상관계수가 음수이면 두 변수 사이에는 음의 상관관계가 존재하며 한 변수가 증가할 때 다른 변수가 감소함을 뜻한다. 이러한 관계는 두 변수에 관한 산점도로 확인할 수도 있다. 음의 상관관계가 강할수록 상관계수는 -1에 가깝고 산점도는 우하향으로 형성되며, 양의 상관관계가 강할수록 상관계수는 1에 가깝고 산점도는 우상향으로 형성된다. 두 변수가 서로 연관되어 있지 않아 상관관계가 없는 경우 상관계수는 0에 가깝다. 상관계수의 값에 따른 상관관계의 정도는 다표 1과 같이 정리할 수 있다.

표 1. 상관관계 해석

상관계수 값	상관관계 해석
$0 \le \|r\| < 0.2$	두 변수 간 상관관계는 매우 약함
$0.2 \le \|r\| < 0.4$	두 변수 간 상관관계는 약함
$0.4 \le \|r\| < 0.6$	두 변수 간 상관관계는 비교적 강함
$0.6 \le \|r\| < 0.8$	두 변수 간 상관관계는 강함
$0.8 \le \|r\| \le 1$	두 변수 간 상관관계는 매우 강함

상관계수는 분석의 대상이 되는 두 변수의 종류에 따라 수식이 다르다. 두 변수가 연속형 변수이면 피어슨 상관계수(Pearson Correlation Coefficient, r_{XY})가 사용된다.

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{13}$$

식 (13)에서 n은 데이터의 수, X_i는 변수 X의 i번째 데이터, $\overline{X}$는 변수 X의 평균을 뜻한다.

두 변수 중 하나의 변수 X가 연속형 변수이고, 나머지 변수 Y가 이진형 변수이면 점이연 상관계수(Point-biserial Correlation Coefficient, r_{pb})를 사용한다.

$$r_{pb} = \frac{M_1 - M_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}}\sqrt{\frac{n_1 n_0}{n^2}} \tag{14}$$

식 (14)에서 M_1은 Y가 1일 때 X의 평균을, M_0은 Y가 0일 때 X의 평균이다.

n_0은 Y가 1인 데이터의 개수, n_1은 Y가 0인 데이터의 개수이다.

상관계수의 통계적 유의성을 확인하는 과정은 귀무가설 "H_0: 두 변수의 상관계수는 0이다(또는 두 변수는 상관관계가 없다).", 연구가설 "H_1: 두 변수의 상관계수는 1이다(또는 두 변수는 상관관계가 있다)."를 수립하는 것에서 출발한다. 수립한 귀무가설은 t-검정으로 검증한다. 일반적인 유의 수준은 0.05로, 식 (15)로 계산한 t 값과 자유도 $(n-2)$를 고려하여 도출한 유의확률(p-value)이 0.05보다 작으면 귀무가설을 기각한다. 본 논문에서는 상관계수는 소수 셋째 자리에서, 유의확률은 소수 넷째 자리에서 반올림하여 계산하였다.

$$T = r\sqrt{\frac{n-2}{1-r^2}} \tag{15}$$

2. 데이터

1) 데이터 수집

코로나19 확산에 관한 데이터는 크게 질병관리청 코로나바이러스감염증-19 공식 홈페이지(http://ncov.mohw.go.kr/), BigKinds(https://www.bigkinds.or.kr/)에서 수집하였다. Bigkinds는 한국언론진흥재단이 국내 주요 54개 언론사에서 보도하는 뉴스 데이터를 종합하여 제공하는 뉴스 빅데이터 시스템이다.

질병관리청 코로나19 홈페이지에서 수집한 데이터는 일간 확진자의 발생 및 사망 현황, 코로나19로 인한 일간 재원 위중증 환자 수, 일간 코로나19 예방접종 현황, 주간 코로나19 변이 검출현황, 사회적 거리두기 현황, 일간 병상가동률 및 재택치료자 현황이다. 수집한 일간 확진자 발생 및 사망 현황 중 확진자 수에 관한 데이터는 크게 발생, 연령대, 성별, 시도로 구분된다. 먼저 확진자의 발생은 국내발생과 해외유입으로 나뉜다. 확진자의 나이는 9개 그룹(0~9세, 10~19세, 20~29세, 30~39세, 40~49세, 50~59세, 60~69세, 70~79세, 80세 이상)으로 분류되며, 확진자의 발생 위치는 전국 17개 시도(서울, 부산, 대구, 인천, 광주, 대전, 울산, 세종, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주)로 구분된다.

코로나19로 인한 재원 위중증 환자 수는 고유량(high flow) 산소요법, 인공호흡기, 체외막산소공급(ECMO: Extracorporeal membrane oxygenation), 지속적인 대체요법(CRRT: Continuous Renal Replacement Therapy) 등으로 격리 치료 중인 환자를 뜻하며(질병관리청, 2021년 11월 16일), 2020년 3월 28일 이후로 일간 데이터가 제공되었다. 질병관리청 코로나19 홈페이지에 따르면 사회적 거리두기는 2020년 6월 28일에 세 단계(1단계, 2단계, 3단계)로 정비된 후, 2020년 11월 7일에는 다섯 단계(1단계, 1.5단계, 2단계, 2.5단계, 3단계)로 개편되었다. 2021년 7월 1일에 네 단계

(1단계, 2단계, 3단계, 4단계)로 조정되었으며 2021년 11월부터 단계적 일상회복 체계로 전환하고 2022년 4월 18일부터 사회적 거리두기 관련 모든 조치는 해제되었다. 따라서 사회적 거리두기 단계가 일정 단계로 유지되더라도 그 강도는 시기에 따라 매우 달랐다. 또한, 사회적 거리두기는 지자체마다 그 강도와 시행 기간의 측면에서 차이가 있으므로, 서로 다른 두 지역의 사회적 거리두기 단계가 같더라도 시행 내용이 서로 다른 경우도 빈번하게 발생하였다.

코로나19 예방접종 현황은 백신 종류와 접종 유형별로 제공되고 백신은 화이자, 모더나, 아스트라제네카, 얀센, 노바백스, 기타, 전체로 구분된다. 백신의 접종 유형은 1차부터 4차까지 구분되며 얀센 백신의 경우 1차와 2차를 합한 데이터가 제공되었다. 코로나19 변이 검출현황 데이터는 WHO에서 코로나19에 대한 우려 변이 바이러스로 지정하였던 알파, 베타, 감마, 델타, 오미크론 변이 각각의 주간 검출 건수이다.

병상가동률과 재택치료자 현황은 2022년 3월 16일 이후로 제공되는 일간 데이터이다. 병상가동률은 중환자 전담치료병상(위중증), 준-중환자 병상(준중증), 감염병 전담병원(중등증), 생활치료센터(경증)에 대해 각 병원의 코로나19 병상이 가용한 정도를 퍼센트(%)로 수치화한 값이다. 재택치료자는 일일 신규 재택치료자와 누적 재택치료자로 구분된다. 중환자 전담치료병상의 경우 집계 시점의 중환자 병상 사용의 현황으로, 재원 위중증 환자 수와 상이할 수 있다.

BigKinds에서 수집한 데이터는 일간 코로나19 확산 관련 뉴스의 수로, 단어 '코로나19', '코로나', '코로나 바이러스', '신종 코로나바이러스', 'COVID-19', '코비드19' 중 한 개 이상을, 단어 '확진자'를 반드시 포함하는 뉴스의 수를 일별로 합산하여 확보하였다. 코로나 검사 수가 주말에는 현상이 있으므로, 어떤 날짜가 특정 요일에 해당하면 1을, 해당하지 않으면 0을 부여하여 데이터를 구성하였다. 예를 들어, 특정 일자의 월요일 여부는 해당 날짜가 월요일인지 아닌지를 나타낸다. 예측 수행을 위해 수집한 모든 데이터는 변수로 처리하였다.

2) 데이터 가공

확진자 발생 현황에서 연령대는 아동기 및 청소년기(20세 미만)와 성년기(20세 이상 60세 미만), 노년기(60세 이상)으로 구분하였다. 시도 구분의 기준이 되는 17개 시도는 수도권(서울, 인천, 경기)과 비수도권(서울, 인천, 경기를 제외한 나머지 14개 시도)으로 재구분하였다. 또한, 일간 확진자 수의 예측을 위해 주간 데이터인 코로나19 변이 검출현황 데이터는 7로 나눈 후 각 주에 해당하는 일자들의 일간 데이터로 변환하여 활용하였다.

본 논문에서는 사회적 거리두기 단계를 전국 단위의 변수로써 사용하기 위해, 전국을 권역 단위(수도권, 충청권, 호남권, 경북권, 경남권, 강원, 제주)로 구분한 후 모든 지자체에서 공통으로 제한하였던 식당·카페 영업 가능 시간과 사적 모임 가능 인원수를 기준으로 사회적 거리두기의 단계를 재구성하였다. 표 2는 재구성한 사회적 거리두기 단계의 상세 기준을 나타내며, 단계는 0부터 5까지의 정수로써 제재의 강도와 비례한다. 단계 0과 1은 식당·카페 영업 가능 시간과 사적 모임 가능 인원수가 제한되지 않았음을 나타낸다는 점에서 공통점이 있으나, 단계 0은 사회적 거리두기의 모든 조치가 해제된 때를, 단계 1은 다른 사회적 거리두기 조치가 시행되던 때를 수치화한 값이다. 사회적 거리두기의 강도는 각 지역의 코로나19 확산 실정에 따라 같은 권역에 속하는 지역일지라도 서로 다를 수 있으나, 사회적 거리두기를 전국적인 변수로 고려하는 본 논문에서는 각 권역에서 가장 널리 적용된 기준을 우선적으로 고려하였다. 또한, 동일한 일자에서 사회적 거리두기 시행 기준이 시간 단위로 달라지면, 각 기준이 시행된 시간대에 따라 사회적 거리두기 단계를 비례배분하여 일간 사회적 거리두기 기준을 계산하였다. 예를 들면 2021년 7월 12일부터 2021년 9월 5일까지, 수도권의 사적모임 가능 인원수는 18시 이전에는 4명, 18시 이후에는 2명이었으므로 시간에 따라 비례배분하여 해당 일자의 사적모임 가능

인원수는 3.5명으로 계산하였다. 권역별로 도출한 사회적 거리두기 단계는 지역의 인구수를 기준으로 비례배분하는 단계를 거친 후, 모든 권역의 수치를 합산함으로써 전국의 일자별 사회적 거리두기 단계로 계산하였다. 표 3은 각 권역에서 식당·카페 영업 가능 시간과 사적 모임 가능 최대 인원수를 제한하였던 시기를 일간을 기준으로 합산한 값을 나타낸다.

표 2. 사회적 거리두기 단계의 재구성

식당·카페 영업 가능 시간	사적 모임 가능 인원수	재구성한 사회적 거리두기 단계
제한 없음 (사회적 거리두기 해제)	제한 없음 (사회적 거리두기 해제)	0
제한 없음	제한 없음	1
21시까지	49명·99명 이내	2
0시까지	4명·6명·8명·10명·12명 이내	3
21시·22시까지	4명·6명·8명 이내	4
21시·22시까지	3.5명·4명 이내	5

표 3. 7개 권역의 사회적 거리두기 시행 기간 (단위: 일)

		사적 모임 가능 최대 인원수 (명)								
		3.5	4	6	8	10	12	49	99	제한 없음 (사회적 거리두기 해제)
식당·카페 영업 가능 최대 시간	21시	14	582	231	0	0	0	30	110	0
	22시	42	729	140	266	0	0	0	0	0
	23시	0	0	112	91	0	0	0	0	0
	24시	0	0	0	0	84	0	0	0	0
	제한없음 (사회적 거리두기 해제)	0	532	30	158	35	210	524	0	1379 (1071)

3. 실험 설계

1) 데이터 확인

각 데이터의 유형과 분포를 확인한 결과 발생별, 연령대별, 지역별 확진자 수, 사망자 수, 입원환자 수, 코로나19 확산 관련 뉴스 검색 수 및 재택치료자 수는 정수형이었으며 변이 검출 수와 재구성한 사회적 거리두기 단계, 병상가동률은 실수형이었다. 예측 대상인 일간 확진자 수는 2020년 2월 23일 이후부터 100을 상회하였고, 결측치가 없었으므로 2020년 2월 23일을 시작으로 분석 기간을 설정하였다.

2) 분석 기간 설정

코로나19 변이 검출이 2020년 12월 28일부터 지속적으로 확인되었다는 점, 오미크론 변이를 제외한 여타 변수의 검출 수가 2022년 3월 20일 이후로는 0을 기록하였다는 점을 고려하여 분석 기간을 세 구간으로 나누어 실험을 진행하였다. 첫 번째 실험(실험 1)은 변이 발생 전 코로나19 확산기, 두 번째 실험(실험 2)은 알파, 베타, 감마, 델타, 오미크론 변이가 창궐하던 코로나19 변이 확산기, 세 번째 실험(실험 3)은 오미크론 변이 확산기에서의 확진자 수를 예측하였다. 각 실험의 시작일과 종료일은 각각 일요일과 토요일로 설정하였으며 이를 요약하면 표 4와 같다. 부록 1은 각 실험에서 가용한 일별 데이터를 나타낸다.

실험 1의 경우, 확진자 수를 예측하는 국내 논문(Bae & Kim, 2021)과 성능을 비교하는 분석을 추가로 진행하였다. 해당 논문은 2020년 6월 4일부터 12월 21일까지, 202일에 이르는 기간에서 발생한 일간 확진자 수를 예측하였고 분석 기간을 7개로 나누어 실험을 7번 진행하였다. 각 실험의 학습 데이터와 검증 데이터, 테스트

데이터를 나타내면 표 5와 같으며, 본 논문에서는 검증 데이터를 별도로 설정하지 않았으므로 검증 데이터와 학습 데이터를 하나의 학습 데이터로써 활용하였다.

표 4. 실험 설명

실험 구분	분석 기간	설명
실험 1	2020.02.23.(일)~2020.12.27.(토) (309일)	코로나19 변이 등장 이전
실험 2	2020.12.28.(일)~2022.03.19.(토) (447일)	코로나19 변이 등장 이후, 오미크론 변이 100% 검출 이전
실험 3	2022.03.20.(일)~2022.09.17.(토) (182일)	오미크론 변이 100% 검출 이후

표 5. 국내논문의 데이터 구분

데이터 구분	실험 1	실험 2	실험 3	실험 4	실험 5	실험 6	실험 7
학습 데이터	2020.01.20.~ 2020.05.24. (126일)	2020.01.20.~ 2020.06.23. (156일)	2020.01.20.~ 2020.07.23. (186일)	2020.01.20.~ 2020.08.21. (215일)	2020.01.20.~ 2020.09.20. (245일)	2020.01.20.~ 2020.10.21. (276일)	2020.01.20.~ 2020.11.20. (306일)
검증 데이터	2020.05.25.~ 2020.06.03. (10일)	2020.06.24.~ 2020.07.03. (10일)	2020.07.24.~ 2020.08.02. (10일)	2020.08.22.~ 2020.08.31. (10일)	2020.09.21.~ 2020.09.30. (10일)	2020.10.22.~ 2020.10.31. (10일)	2020.11.21.~ 2020.11.30. (10일)
테스트 데이터	2020.06.04.~ 2020.07.03. (30일)	2020.07.04.~ 2020.08.02. (30일)	2020.08.03.~ 2020.09.01. (30일)	2020.09.01.~ 2020.09.30. (30일)	2020.10.01.~ 2020.10.31. (31일)	2020.11.01.~2020. 11.30. (30일)	2020.12.01.~ 2020.12.21. (21일)

3) 입력변수 선정

코로나19 잠복기가 최대 14일이므로(Central Disease Control Headquarters & Central Disaster Management Headquarters, 2021) 특정 과거 시점으로부터 최소 1일 전부터 최대 14일 전까지의 확진자 수를 모든 실험에서 개별 입력변수로 고려하였다. 실험은 사용하는 입력변수의 종류에 따라 구분된다. 먼저, 각 실험에서 가용한 변수에 대해 $(t-1)$ 시점의 데이터와 t 시점의 확진자 수와의 상관계수를 구한 후 상관계수의 유의확률이 0.05 이하인 변수만을 입력변수로 활용하였다. 이때 모든 입력변수를 활용한 예측을, 확진자 수와의 상관계수가 0.9보다 큰 입력변수, 상관계수가 0.8보다 큰 입력변수, 0.7보다 큰 입력변수, 0.6보다 큰 입력변수를 사용한 예측과 비교하여 어떤 변수를 사용해야 성능이 향상되는지를 확인하였다. 모든 실험은 각각 5회 반복실험한 후, 각 실험에서 도출한 예측의 평균값을 비교하여 어떤 입력변수를 사용할 때 더 좋은 예측을 수행하는지와, 동일한 분석 기간에서는 어떤 모델의 예측이 더 우수한지를 확인하였다.

모든 모델의 학습 데이터와 테스트 데이터의 비율은 전체 데이터의 90%, 10%로 설정하였다. 모델의 성능 평가 측도에는 평균 제곱근 편차(RMSE: Root Mean Square Error)가 대표적이나(식 (16)), 데이터 단위에 영향을 받아 큰 값을 예측할수록 RMSE의 값이 커진다는 단점을 보완하기 위해 평균 절대 백분율 오차(MAPE: Mean Absolute Percentage Error)를 채택하여 예측 기간이 다른 모델들 간 비교를 수행하였다(식 (17)). 식 (16), 식 (17)에서 n은 데이터 수, y는 실제값, $\hat{y}$는 예측값이며, 본 논문에서 MAPE는 소수 셋째 자리에서 반올림하여 나타내었다.

$$RMSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n} \tag{16}$$

$$MAPE(\%) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \tag{17}$$

4. 실험 1: 변이 발생 전 코로나19 확산기

변이 발생 전 코로나19 확산기에서 입력변수로 고려한 변수는 해외유입, 연령대별 확진자 수, 수도권·비수도권 확진자 수, 코로나19로 인한 사망자 수 및 재원 위중증 환자 수, 재구성한 사회적 거리두기 단계, 코로나19 확산 관련 뉴스 기사 수, 과거 확진자 수이다. 이 중 특정 요일 여부에 관한 변수를 제외한 모든 변수가 확진자 수와 통계적으로 유의한 양의 상관관계가 있는 것으로 확인되었다. 실험 1에서 확진자 수와 가장 큰 연관이 있는 변수는 전날 발생한 확진자 수로 나타났으며, 확진자 수와의 상관계수가 0.9 이상인 변수는 8개, 0.8 이상인 변수는 12개, 0.7 이상인 변수는 17개, 0.6 이상인 변수는 21개이다. 실험 1에서 입력변수로 고려한 모든 변수와 확진자 수와의 상관분석 결과는 표 6으로 요약된다.

표 6. 실험 1에서 고려한 입력변수

상관계수 구분	변수	상관계수	유의확률
0.9 이상	1일 전 확진자 수	0.97	0.000
	성년기 확진자 수	0.96	0.000
	2일 전 확진자 수	0.95	0.000
	아동기 및 청소년기 확진자 수	0.94	0.000
	노년기 확진자 수	0.94	0.000
	3일 전 확진자 수	0.94	0.000
	4일 전 확진자 수	0.92	0.000
	5일 전 확진자 수	0.90	0.000
0.8 이상 0.9 미만	수도권 확진자 수	0.89	0.000
	6일 전 확진자 수	0.88	0.000
	7일 전 확진자 수	0.86	0.000
	8일 전 확진자 수	0.83	0.000

상관계수 구분	변수	상관계수	유의확률
0.7 이상 0.8 미만	9일 전 확진자 수	0.79	0.000
	10일 전 확진자 수	0.76	0.000
	비수도권 확진자 수	0.74	0.000
	11일 전 확진자 수	0.73	0.000
	12일 전 확진자 수	0.70	0.000
0.6 이상 0.7 미만	13일 전 확진자 수	0.68	0.000
	사망자 수	0.68	0.000
	14일 전 확진자 수	0.65	0.000
	재원 위중증 환자 수	0.62	0.000
0.6 미만	사회적 거리두기 단계 (재구성)	0.49	0.000
	코로나19 확산 관련 뉴스 기사 수	0.33	0.000
	해외유입	0.19	0.001

5. 실험 2: 코로나19 변이 확산기

2020년 12월 28일, 국내 최초의 코로나19 변이인 알파 변이가 처음 검출된 이후 베타 변이, 감마 변이, 델타 변이, 오미크론 변이가 차례로 확산하였다. 실험 2의 분석 기간은 모든 변이가 검출되던 시기이자, 백신 접종이 시작되고 사회적 거리두기의 강도가 다른 기간보다 높게 책정되는 등 정부개입이 가장 활성화된 기간이다. 이에 따라 백신 보급과 같이 실험 1에서 확보할 수 없었던 데이터가 새롭게 등장하며 입수 가능한 데이터가 늘어나므로, 실험 1에서 활용한 입력변수를 실험 2의 예측에 그대로 활용한다면 예측의 정확도가 저하될 수 있다고 판단하였다. 이를 확인하기 위해 실험 2에서는 실험 1에서 고려한 변수, 코로나19 변이 검출 수, 백신 접종 수를 실험 2의 독자적 입력변수로 설정하여 예측을 수행하였고, 실험 1의 입력변수를 사용하였을 때의 예측과 비교하였다.

실험 2의 분석 기간에서 다수 백신의 접종은 1차부터 4차까지 이루어졌다. 이때 아스트라제네카 백신과 기타 백신은 3차 접종까지 보급되었고 화이자, 모더나, 노바백스 백신의 4차 접종은 2022년 2월 17일 이후 진행되었다는 점을 고려하여, 실험 2에서는 1차 접종부터 3차 접종까지 예측에 고려하였다. 실험 2의 독자적 입력변수와 확진자 수를 상관분석한 결과, 코로나19 확산 관련 뉴스 기사 수, 특정 요일 여부에 관한 변수, 일부 백신 접종 수를 제외한 모든 변수가 확진자 수와 통계적으로 유의한 상관관계가 있는 것으로 확인되었다. 실험 2에서 확진자 수와 가장 큰 연관이 있는 변수는 8일 전 확진자 수로 나타났다. 확진자 수와의 상관계수가 0.9 이상인 변수는 19개, 0.8 이상인 변수는 20개, 0.7 이상인 변수는 22개, 0.6 이상인 변수는 24개이다. 실험 2에서 입력변수로 고려한 모든 변수와 확진자 수와의 상관분석 결과는 표 7로 요약된다.

표 7. 실험 2에서 고려한 입력변수

상관계수 구분	변수	상관계수	유의확률
0.9 이상	8일 전 확진자 수	0.98	0.000
	7일 전 확진자 수	0.98	0.000
	5일 전 확진자 수	0.97	0.000
	13일 전 확진자 수	0.97	0.000
	아동기 및 청소년기 확진자 수	0.97	0.000
	비수도권 확진자 수	0.97	0.000
	4일 전 확진자 수	0.97	0.000
	14일 전 확진자 수	0.97	0.000
	12일 전 확진자 수	0.97	0.000
	6일 전 확진자 수	0.97	0.000
	11일 전 확진자 수	0.96	0.000
	3일 전 확진자 수	0.96	0.000
	성년기 확진자 수	0.96	0.000
	1일 전 확진자 수	0.96	0.000
	수도권 확진자 수	0.96	0.000
	10일 전 확진자 수	0.95	0.000
	2일 전 확진자 수	0.95	0.000
	9일 전 확진자 수	0.95	0.000
	노년기 확진자 수	0.94	0.000
0.8 이상 0.9 미만	사망자 수	0.86	0.000
0.7 이상 0.8 미만	노바백스 백신 3차 접종 수	0.75	0.000
	노바백스 백신 2차 접종 수	0.71	0.000
0.6 이상 0.7 미만	노바백스 백신 1차 접종 수	0.65	0.000
	오미크론 변이 검출 수	0.64	0.000

상관계수 구분	변수	상관계수	유의확률
0.6 미만	재원 위중증 환자 수	0.51	0.000
	해외유입	0.14	0.002
	사회적 거리두기 단계 (재구성)	0.11	0.017
	코로나19 확산 관련 뉴스 기사 수	0.10	0.037
	감마 변이 검출 수	-0.10	0.028
	화이자 백신 2차 접종 수	-0.13	0.006
	화이자 백신 1차 접종 수	-0.14	0.003
	전체 백신 2차 접종 수	-0.14	0.003
	전체 백신 1차 접종 수	-0.15	0.002
	베타 변이 검출 수	-0.15	0.002
	기타 백신 1차 접종 수	-0.16	0.001
	기타 백신 2차 접종 수	-0.18	0.000
	알파 변이 검출 수	-0.19	0.000
	델타 변이 검출 수	-0.20	0.000

6. 실험 3: 오미크론 변이 확산기

2020년 12월 28일에 국내에서 처음 검출된 오미크론 변이는 확산세가 다른 변이들보다 증가하여, 2022년 3월 20일 이후로는 모든 확진자가 오미크론 변이에 감염된 것이 확인되었다. 실험 3의 분석 기간에서 백신 접종 수의 증가율은 2022년 4월 18일부터 모든 사회적 거리두기가 해제되며 비교적 낮게 나타났다. 또한, 아스트라제네카 백신의 1차 접종 수는 기록되지 않았으며 질병관리청이 보도자료를 통해 중환자 전담치료병상, 준-중환자 병상, 감염병 전담병원, 생활치료센터 각각의 가동률과 더불어 재택치료자 수를 일별로 누적값과 일별 신규 값을 제공하기 시작하는 등 실험 3의 분석 기간부터 데이터 변화 양상과 데이터 가용성 여부에 변화가 있었다. 그러므로 실험 2에서와 마찬가지로, 시기적으로 앞선 실험인 실험 2에서의 입력변수를 사용하였을 때의 예측과 실험 3의 분석 기간에서 입수 가능한 입력변수를 사용하였을 때의 예측을 비교하였다. 실험 3에서는 실험 2에서 고려한 변수에서 코로나19 변이 검출 수, 아스트라제네카 백신 1차 접종 수를 제외하고 새롭게 입수한 6가지 데이터를 모델의 입력변수로 고려하였다. 이 중 해외유입, 코로나19 확산 관련 뉴스 기사 수, 월요일을 제외한 특정 요일 여부에 관한 변수와 일부 백신 접종 수를 제외한 모든 변수가 확진자 수와 통계적으로 유의한 상관관계가 있는 것으로 확인되었다. 실험 3에서 확진자 수와 가장 큰 연관이 있는 변수는 7일 전 확진자 수로 나타났고, 확진자 수와의 상관계수가 0.9 이상인 변수는 7개, 0.8 이상인 변수는 25개, 0.7 이상인 변수는 30개, 0.6 이상인 변수는 33개이다. 실험 3에서 입력변수로 고려한 모든 변수와 확진자 수와의 상관분석 결과는 표 8로 요약된다.

표 8. 실험 3에서 고려한 입력변수

상관계수 구분	변수	상관계수	유의확률

상관계수 구분	변수	상관계수	유의확률
0.9 이상	7일 전 확진자 수	0.95	0.000
	아동기 및 청소년기 확진자 수	0.92	0.000
	6일 전 확진자 수	0.92	0.000
	1일 전 확진자 수	0.91	0.000
	성년기 확진자 수	0.91	0.000
	수도권 확진자 수	0.91	0.000
	비수도권 확진자 수	0.91	0.000
0.8 이상 0.9 미만	누적 재택치료자 수	0.88	0.000
	8일 전 확진자 수	0.87	0.000
	신규 재택치료자 수	0.87	0.000
	노년기 확진자 수	0.87	0.000
	5일 전 확진자 수	0.86	0.000
	3일 전 확진자 수	0.86	0.000
	2일 전 확진자 수	0.86	0.000
	4일 전 확진자 수	0.85	0.000
	14일 전 확진자 수	0.84	0.000
	10일 전 확진자 수	0.83	0.000
	13일 전 확진자 수	0.83	0.000
	9일 전 확진자 수	0.83	0.000
	사회적 거리두기 단계 (재구성)	0.81	0.000
	중환자 전담치료병상 가동률	0.81	0.000
	11일 전 확진자 수	0.81	0.000
	재원 위중증 환자 수	0.80	0.000
	12일 전 확진자 수	0.80	0.000
	사망자 수	0.80	0.000
0.7 이상 0.8 미만	노바백스 백신 1차 접종 수	0.77	0.000
	모더나 백신 3차 접종 수	0.76	0.000
	전체 백신 3차 접종 수	0.75	0.000
	화이자 백신 3차 접종 수	0.75	0.000
	준-중환자 병상 가동률	0.75	0.000

상관계수 구분	변수	상관계수	유의확률
0.6 이상 0.7 미만	얀센 백신 3차 접종 수	0.70	0.000
	감염병 전담병원 가동률	0.68	0.000
	모더나 백신 1차 접종 수	0.64	0.000
0.6 미만	코로나19 확산 관련 뉴스 기사 수	0.59	0.000
	노바백스 백신 2차 접종 수	0.59	0.000
	생활치료센터 가동률	0.59	0.000
	모더나 백신 2차 접종 수	0.52	0.000
	아스트라제네카 백신 3차 접종 수	0.50	0.000
	전체 백신 1차 접종 수	0.48	0.000
	화이자 백신 1차 접종 수	0.34	0.000
	전체 백신 2차 접종 수	0.28	0.000
	노바백스 백신 3차 접종 수	0.28	0.000
	월요일 여부	-0.17	0.023
	기타 백신 3차 접종 수	-0.17	0.021
	기타 백신 2차 접종 수	-0.23	0.001

제4장 연구 결과

1. 실험 1

1) 다른 논문과의 성능 비교

본 논문에서 활용한 방법론의 성능을 확인하기 위해 우리나라 확진자 수를 예측하는 다른 논문(Bae & Kim, 2021)과 비교하였다. Bae & Kim (2021)에서 RMSE를 활용하여 예측을 평가하였으므로 본 논문이 수행한 예측 결과 역시 RMSE를 기준으로 비교하였다. 각 모델의 성능은 실험마다 다른 것으로 나타났으며, 시간이 지날수록 RMSE의 값이 커지는 경향이 있었다. 본 논문의 방법론을 Bae & Kim (2021)이 설정한 7개 기간에 적용한 결과, 실험 1, 2에서는 2개의 GRU를 활용한 모델, 실험 3, 4, 6, 7에서는 1개의 RNN을 활용한 모델, 실험 5에서는 3개의 GRU를 활용한 모델이 가장 우수한 예측을 수행하였다. 7개 기간에 대해 본 논문에서 제시한 모든 예측의 평균 RMSE는 80.10인 반면 Bae & Kim (2021)의 평균 RMSE는 132.79로, 본 논문이 상대적으로 정확한 예측을 수행함을 알 수 있으며 각 모델의 평균 RMSE는 국내 논문에서 제시한 RMSE보다 모두 우수하였다. 확진자 수가 200명 미만을 기록하던 기간을 대상으로 하는 실험 1에서 실험 5까지의 실험에서는 본 논문의 예측 성능이 두드러지지 않았으나 확진자 수가 1,000명 이상으로 급격히 증가하는 11월 중순 이후를 포함하는 기간인 실험 6과 실험 7에서의 예측에서는 본 논문의 예측 성능이 훨씬 우수하였다. 특히 본 논문의 RMSE는 Bae & Kim (2021)의 결과보다 최대 439.44가 작게 나타났다. 표 9는 7개 실험에 대해 논문에서 사용한 방법론의 성능을 구체적으로 제시하였다.

표 9. 각 논문의 성능 비교

구분	모델 구분	실험 1	실험 2	실험 3	실험 4	실험 5	실험 6	실험 7	평균
본 논문	RNN 1개	18.46	22.16	92.51	46.95	33.97	75.28	155.41	63.54
	RNN 2개	17.18	21.10	103.36	56.16	34.27	78.15	170.74	68.71
	RNN 3개	20.37	21.21	97.36	58.25	35.87	86.28	171.20	70.08
	LSTM 1개	18.37	21.71	109.82	77.02	36.70	122.09	219.43	86.45
	LSTM 2개	19.37	23.64	121.10	77.06	35.36	117.23	214.34	86.87
	LSTM 3개	19.84	26.14	155.99	83.01	42.73	163.15	362.62	121.93
	GRU 1개	18.01	23.73	105.29	51.89	38.79	89.69	186.17	73.37
	GRU 2개	15.85	20.45	103.91	56.23	36.12	91.04	182.85	72.35
	GRU 3개	19.15	20.78	103.76	59.15	32.14	97.26	210.91	77.59
Bae & Kim, 2021	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

2) 예측 결과

실험 1에서는 2020년 11월 27일부터 12월 27일까지, 31일간 발생한 일별 확진자 수를 예측하였다. 먼저 확진자 수와의 상관계수가 0이 아닐 확률이 0.05 이하로, 통계적으로 유의한 모든 변수를 입력변수로 투입하여 예측을 수행한 결과, 표 10에서 보듯이 최소 MAPE는 9.43%로 계산되었다. 이를 도출한 모델은 14일간의 과거 데이터를 예측에 활용하였고, 은닉층은 1개의 GRU로 구성되었다. 다음으로 확진자 수와의 상관계수가 0.6 이상, 0.7 이상, 0.8 이상, 0.9 이상으로써 확진자 수와 상관관계가 높은 변수만으로 예측을 수행한 결과, 최소 MAPE는 각각 9.58%, 8.40%, 9.01%, 8.71%로 계산되었다. 실험 1에서 구한 최적 모델의 상관계수, 사용된 입력변수의 개수, 최적의 모델, 은닉층 개수, Time Step, MAPE 등은 다음 표와 같다.

표 10. 실험 1에서 구한 최적 모델의 하이퍼파라미터

상관계수	입력변수 개수	최적 모델의 은닉층 종류	최적 모델의 은닉층 개수	Time Step	MAPE
0이 아님	24	GRU	1	14	9.43%
0.6 이상	21	GRU	3	1	9.58%
0.7 이상	17	GRU	3	12	8.40%
0.8 이상	12	LSTM	1	7	9.01%
0.9 이상	8	LSTM	1	9	8.71%

MAPE가 8.40%로 가장 좋은 예측 성능을 보인 최적 모델의 은닉층은 3개의 GRU로 구성되었으며, 12일간의 과거 데이터를 예측에 활용하였다. 이러한 결과를 통해 1~12일 전 확진자 수, 연령대별 확진자 수, 지역별 확진자 수를 활용할 때 예측 성능이 가장 좋고, 여타 변수는 예측 성능을 개선하는 데 기여하지 않음을 확인하였다. 실제 확진자 수와 가장 우수한 예측 모델의 예측값을 비교하면 그림 7과 같다. 그림 7을 살펴보면 실제 확진자 수가 큰 변화가 없을 때는 예측이 상대적으

로 정확하지만, 실제 확진자 수가 급증하거나 급감하는 경우 예측이 하루 이전의 값에 더 가까운 경향을 보이고 있다. 이는 사용하는 데이터가 모두 일간 데이터인 경우 모델이 확진자 수의 급격한 추세 변화에 대응하기에는 데이터의 수집 최소 단위인 하루 이상의 시차가 발생하기 때문인 것으로 해석할 수 있다.

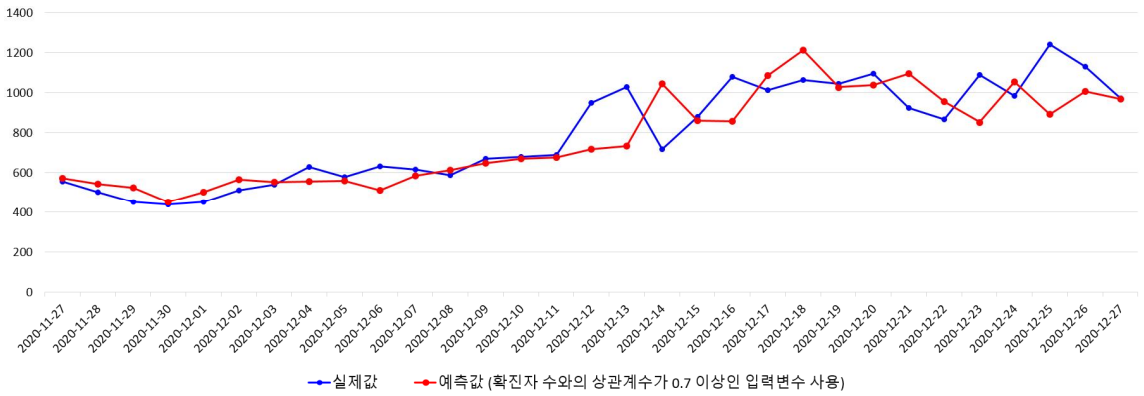

그림 7. 실험 1에서 가장 우수한 예측

2. 실험 2

실험 2에서는 2022년 2월 3일부터 3월 19일까지, 45일간 발생한 일별 확진자 수를 예측하였다. 먼저 확진자 수와의 상관계수가 0이 아닌 모든 변수를 입력변수로 투입하여 예측을 수행한 결과, 최소 MAPE는 12.22%로 계산되었다. 이를 도출한 모델의 은닉층은 1개의 RNN으로 구성되고, 모델은 9일간의 과거 데이터를 예측에 활용하였다. 다음으로 확진자 수와의 상관계수가 0.6, 0.7, 0.8, 0.9 이상인 변수를 사용한 경우 최소 MAPE는 각각 10.76%, 11.10%, 11.12%, 10.65%로 계산되었다. 실험 2에서 구한 최적 모델의 상관계수, 사용된 입력변수의 개수, 최적의 모델, 은닉층 개수, Time Step, MAPE 등은 표 11과 같다.

표 11. 실험 2에서 구한 최적 모델의 하이퍼파라미터

상관계수	입력변수 개수	최적 모델의 은닉층 종류	최적 모델의 은닉층 개수	Time Step	MAPE
0이 아님	38	RNN	1	9	12.22%
0.6 이상	24	GRU	1	13	10.76%
0.7 이상	22	GRU	1	2	11.10%
0.8 이상	20	RNN	3	2	11.12%
0.9 이상	19	GRU	2	14	10.65%

MAPE가 10.65%로 가장 좋은 예측 성능을 보인 최적 모델의 은닉층은 2개의 GRU로 구성되었고 14일간의 과거 데이터를 예측에 활용하였다. 이를 통해 1~14일 전 확진자 수, 연령대별 확진자 수, 지역별 확진자 수를 제외한 모든 변수는 예측 성능을 크게 개선하지 않는 요인으로 확인하였다. 그림 8은 서로 다른 입력변수를 사용한 예측 중 가장 우수한 예측을 나타낸다.

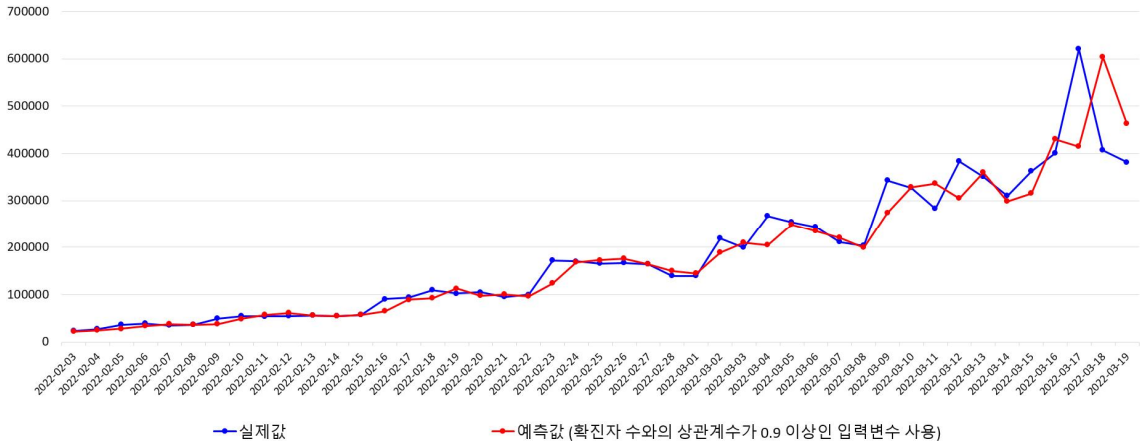

그림 8. 실험 2에서 가장 우수한 예측

3. 실험 3

실험 3에서는 2022년 8월 30일부터 9월 17일까지, 19일간 발생한 일별 확진자 수를 예측하였다. 먼저 확진자 수와의 상관계수가 0이 아닌 모든 변수를 입력변수로 투입하여 예측을 수행한 결과, 최소 MAPE는 13.16%로 계산되었다. 이를 도출한 모델의 은닉층은 1개의 RNN으로 구성되고, 모델은 6일간의 과거 데이터를 예측에 활용하였다. 확진자 수와의 상관계수가 0.6, 0.7, 0.8, 0.9 이상인 변수를 차례로 입력변수로 고려할 때, 최소 MAPE는 각각 14.87%, 15.15%, 13.77%, 12.95%로 계산되었다. 실험 3에서 구한 최적 모델의 상관계수, 사용된 입력변수의 개수, 최적의 모델, 은닉층 개수, Time Step, MAPE 등은 표 12와 같다.

표 12. 실험 3에서 구한 최적 모델의 하이퍼파라미터

상관계수	입력변수 개수	최적 모델의 은닉층 종류	최적 모델의 은닉층 개수	Time Step	MAPE
0이 아님	61	RNN	1	6	13.16%
0.6 이상	33	RNN	3	1	14.87%
0.7 이상	30	RNN	2	1	15.15%
0.8 이상	25	LSTM	1	1	13.77%
0.9 이상	7	GRU	1	11	12.95%

MAPE가 12.95%로 가장 좋은 예측 성능을 보인 최적 모델의 은닉층은 1개의 GRU로 구성되었으며, 11일간의 과거 데이터를 예측에 활용하였다. 이를 통해 1일, 6일, 7일 전 확진자 수, 아동기 및 청소년기 확진자 수, 성년기 확진자 수, 지역별 확진자 수를 제외한 모든 변수는 예측 성능을 크게 개선하지 않는 요인으로 확인하였다. 그림 9는 서로 다른 입력변수를 사용한 예측 중 가장 우수한 예측을 나타낸다.

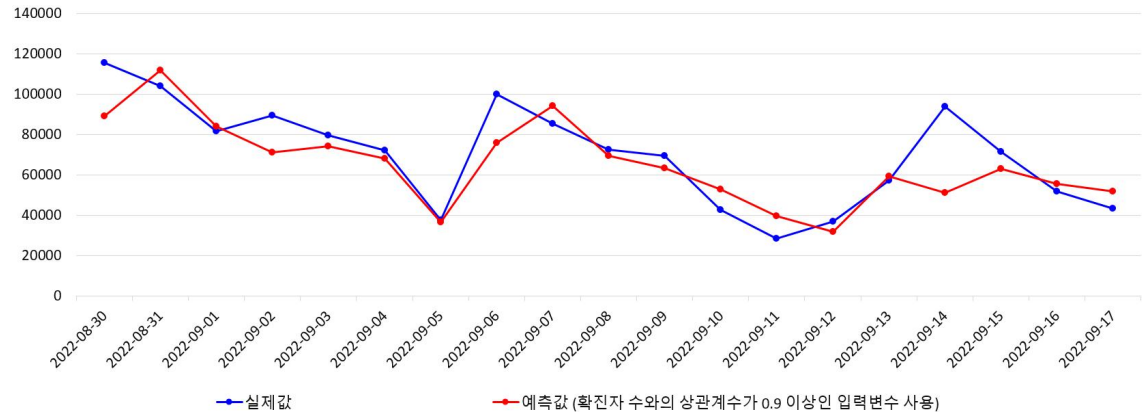

그림 9. 실험 3에서 가장 우수한 예측

제5장 결 론

1. 연구 의의 및 제언

본 논문은 딥러닝 기법과 공공데이터를 사용하여 국내에서 발생하는 확진자 수를 예측하였다. 기존 연구들이 제시하였던 머신러닝 및 딥러닝 기반의 예측은 코로나19 확산 방지를 위한 정부의 방역 정책이 코로나19 확산에 영향을 미칠 수 있음을 간과하였다. 또한, 기존 연구들은 한정적인 데이터만을 사용하였고 코로나19 확산 양상이 크게 변하지 않았던 초기에 한정하였으며, 코로나19 확산 양상에 따라 변수의 중요도가 달라질 수 있음을 고려하지 않아 예측이 부정확하였다는 한계가 있었다. 본 논문은 전국의 지역과 시간에 따라 다른 사회적 거리두기 단계를 일관된 기준으로 재구성하여 사람들의 실제 생활에 직접적인 영향을 미치는 정책을 예측에 고려하였고, 코로나19 변이 검출 수, 백신 접종 수 등 코로나19 확산과 관련하여 다양한 데이터를 단계마다 체계적으로 사용하였다. 그 결과, 모든 실험에서 과거의 확진자 수, 연령대별 및 지역별 확진자 수를 활용할 때 예측 성능이 가장 우수하였다. 특히, 본 논문은 확진자 수 예측에서 중요하게 작용하는 입력변수와 최적 모델, 예측에 반영할 과거 데이터의 시점 수는 코로나19 변이 발생 및 확산세에 따라 상이하다는 점에 주목하였다. 변이 등장과 같이 코로나19 확산에 중대한 영향을 미치는 사건이 발생하였을 때, 분석 기간에 따라 입력변수를 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 모델의

단계적인 개발이 효과적일 수 있다는 시사점을 준다. 본 논문에서 제안하는 코로나19 관련 정책의 개선 방향은 다음과 같다. 먼저, 모델에 투입할 변수를 선별하기 위한 상관분석에서, 오미크론 변이가 100% 검출된 이후 월요일 여부를 나타내는 이진변수가 확진자 수와 유의한 음의 상관관계가 있는 점은 일요일에 발생한 확진자 수가 다른 요일에서 발생한 확진자 수보다 감소하였음을 뜻한다. 이는 각 일자의 데이터가 0시를 기준으로 집계되었으며, 일요일에는 상대적으로 검사 수가 적기 때문이라고 해석하였다. 코로나19에 감염되었으나 양성 판정을 받지 않아 자가격리 안내를 수신하지 못한 사람들이 다른 사람들과 접촉하여 코로나19 확산을 가속할 수 있고, 다수의 코로나19 임시선별검사소 및 선별진료소는 일요일에 운영하지 않아 사람들이 코로나19 검사에 적극적으로 참여하기 어려울 수 있다. 이러한 문제는 사람들에게 코로나19 증상을 소개하고 코로나19 확산에 대한 경각심을 일깨우는 공익광고의 노출수를 일요일에 늘려 검사를 독려하거나, 코로나19 검사 관련 의료 시설에 대한 지원을 강화함으로써 완화할 수 있을 것이다.

본 논문에서 제안하는 데이터 관리 측면의 정책 개선 방향은 다음과 같다. 본 논문은 2022년 11월에 입수한 데이터를 근거로 하여, 당일 0시까지 집계된 데이터를 활용할 수 있었으나 코로나19 관련 데이터의 책정 기준은 수시로 변화하였다. 일례로 데이터의 신속한 전달을 위해, 2020년 2월 2일부터 3월 1일까지 코로나19 국내 발생 현황은 8시, 9시, 16시, 19시를 기준으로 집계되는 등 제공 시점이 다양하여, 본 논문에서 활용한 데이터와 차이가 있었다. 또한, 데이터의 단위 또한 일간과 주간 등으로 일정하지 않았으므로 본 논문에서는 주간 데이터를 일간 데이터 예측에 활용하기 위해 주간 데이터를 7로 나눈 값을 일간 데이터로 활용하였다. 이로 인해 주간 데이터가 일간 데이터가 갖는 변동을 충분히 설명하지 못하여 예측의 정확도가 일간 데이터만을 사용할 때보다 우수하지 않을 수 있다. 따라서 데이터 정책을 강화하여 일관된 주기에 따라 데이터를 산정, 안정적으로 확보할 수 있게 한다면 코로나19 예측 성능을 높일 수 있을 것이다. 특히, 주간 데이터 중 오미크론

변이 검출 수는 확진자 수와 강한 양의 상관관계를 가지므로 변이 검출에 관한 데이터가 일간으로 주어진다면 예측의 정확도를 높일 수 있을 것으로 예상하고, 모델의 신뢰성 또한 높일 수 있을 것이다.

2. 연구 한계 및 향후 개선 방향

본 논문은 크게 두 가지 측면에서 개선될 수 있다. 첫째로, 사회적 거리두기 변수의 책정 기준을 보완하고 다른 방역 정책을 예측에 반영할 수 있다. 본 논문에서는 전국 단위의 식당·카페 영업 가능 시간과 사적 모임 가능 인원만을 고려하였으나, 사적 모임보다 규모가 큰 행사나 집회, 혹은 다중이용시설의 단계별 운영 제한 및 유형별 또한 참고할 수 있다. 사회적 거리두기 이외에도 마스크 착용 의무화, 공적 마스크 제도, 백신패스, 자가격리 의무화 등 다양한 정책 현황을 예측의 입력변수로 사용할 수 있을 것이다. 검역 및 입국 규제와 같이 다른 국가와의 교류를 제재하는 정책의 경우, 국내 정책뿐만 아니라 해외에서 우리나라에 대해 시행하는 정책도 추가로 고려할 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. 데이터 발표 기관 및 집계 시점의 차이로 인해 변동이 있는 데이터는 최신의 자료를 사용하여 데이터의 안정성을 고려한 것처럼, 향후 연구에서도 안정된 데이터를 확보해야 한다. 이동 및 의료 데이터와 같이 민감 데이터에 개인정보 비식별 조처를 해 예측에 사용할 수도 있을 것이다. 이처럼 사용할 입력 데이터를 충분히 확보한다면, 예측이 필요한 확진자 수의 변동 폭을 잘 포착하는 향상된 모델을 차후 구축할 수 있을 것이다. 또한, 지역별 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 방지대책에도 일조할 수 있을 것이다.

참 고 문 헌

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.

COVID-19 ForecastHub (2021, November), Forecast Evaluations, Retrieved

November 12, https://covid19forecasthub.org/eval-reports/?state=US&week=2021-11-12#Incident_Case_Forecasts_(county).

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.

Morel, J. D., Morel, J. M. and Alvarez, L. (2022) Learning from the past: a short term forecast method for the COVID-19 incidence curve, medRxiv preprint, 1-19.

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&Me

tric=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.

부 록

<부록 1> 실험 별 가용한 일별 데이터

각 실험에서 수집할 수 있는 데이터는 다음과 같이 요약된다.

실험 1, 실험 2, 실험 3에서 가용한 일별 데이터

변수	실험 1	실험 2	실험 3
해외유입	O	O	O
연령대별, 지역별 확진자 수	O	O	O
사망자 수	O	O	O
재원 위중증 환자 수	O	O	O
사회적 거리두기 단계 (재구성)	O	O	O
코로나19 확산 관련 뉴스 기사 수	O	O	O
특정 요일 여부	O	O	O
1~14일 전 확진자 수	O	O	O
전체 백신 1~4차 접종 수	X	O	O
화이자 백신 1~4차 접종 수	X	O	O
모더나 백신 1~4차 접종 수	X	O	O
아스트라제네카 백신 1차 접종 수	X	O	X
아스트라제네카 백신 2차 접종 수	X	O	O
아스트라제네카 백신 3차 접종 수	X	O	O
아스트라제네카 백신 4차 접종 수	X	X	O
얀센 백신 1차, 2차 접종 수	X	O	O
얀센 백신 3차 접종 수	X	O	O
얀센 백신 4차 접종 수	X	O	O
노바백스 백신 1~4차 접종 수	X	O	O
기타 백신 1차 접종 수	X	O	O
기타 백신 2차 접종 수	X	O	O

변수	실험 1	실험 2	실험 3
기타 백신 3차 접종 수	X	O	O
기타 백신 4차 접종 수	X	X	O
변이 검출 수 (알파, 베타, 감마, 델타, 오미크론)	X	O	X
중환자 전담치료병상 가동률	X	X	O
준-중환자 병상 가동률	X	X	O
감염병 전담병원 가동률	X	X	O
생활치료센터 가동률	X	X	O
누적 재택치료자 수	X	X	O
신규 재택치료자 수	X	X	O

<부록 2> 각 실험에서 고려한 변수

실험 1, 2, 3에서 수집한 모든 변수에 대해, 확진자 수와 상관관계를 구하면 다음과 같이 요약된다.

실험 1에서 고려한 변수

변수	상관계수	유의확률
해외유입	0.19	0.001
아동기 및 청소년기 확진자 수	0.94	0.000
성년기 확진자 수	0.96	0.000
노년기 확진자 수	0.94	0.000
수도권 확진자 수	0.89	0.000
비수도권 확진자 수	0.74	0.000
사망자 수	0.68	0.000
재원 위중증 환자 수	0.62	0.000
사회적 거리두기 단계 (재구성)	0.49	0.000
코로나19 확산 관련 뉴스 기사 수	0.33	0.000
월요일 여부	-0.05	0.397
화요일 여부	-0.05	0.373
수요일 여부	0.00	0.977
목요일 여부	0.01	0.866
금요일 여부	0.02	0.696
토요일 여부	0.04	0.456
일요일 여부	0.02	0.689
1일 전 확진자 수	0.97	0.000
2일 전 확진자 수	0.95	0.000
3일 전 확진자 수	0.94	0.000
4일 전 확진자 수	0.92	0.000
5일 전 확진자 수	0.90	0.000
6일 전 확진자 수	0.88	0.000
7일 전 확진자 수	0.86	0.000

변수	상관계수	유의확률
8일 전 확진자 수	0.83	0.000
9일 전 확진자 수	0.79	0.000
10일 전 확진자 수	0.76	0.000
11일 전 확진자 수	0.73	0.000
12일 전 확진자 수	0.70	0.000
13일 전 확진자 수	0.68	0.000
14일 전 확진자 수	0.65	0.000

실험 2에서 고려한 변수

변수	상관계수	유의확률
해외유입	0.14	0.002
아동기 및 청소년기 확진자 수	0.97	0.000
성년기 확진자 수	0.96	0.000
노년기 확진자 수	0.94	0.000
수도권 확진자 수	0.96	0.000
비수도권 확진자 수	0.97	0.000
사망자 수	0.86	0.000
재원 위중증 환자 수	0.51	0.000
사회적 거리두기 단계 (재구성)	0.11	0.017
코로나19 확산 관련 뉴스 기사 수	0.10	0.037
월요일 여부	-0.03	0.508
화요일 여부	-0.03	0.579
수요일 여부	0.01	0.800
목요일 여부	0.03	0.527
금요일 여부	0.01	0.772
토요일 여부	0.02	0.675
일요일 여부	-0.02	0.703
1일 전 확진자 수	0.96	0.000
2일 전 확진자 수	0.95	0.000
3일 전 확진자 수	0.96	0.000
4일 전 확진자 수	0.97	0.000
5일 전 확진자 수	0.97	0.000
6일 전 확진자 수	0.97	0.000
7일 전 확진자 수	0.98	0.000
8일 전 확진자 수	0.98	0.000
9일 전 확진자 수	0.95	0.000
10일 전 확진자 수	0.95	0.000
11일 전 확진자 수	0.96	0.000
12일 전 확진자 수	0.97	0.000
13일 전 확진자 수	0.97	0.000
14일 전 확진자 수	0.97	0.000

변수	상관계수	유의확률
전체 백신 1차 접종 수	-0.15	0.002
전체 백신 2차 접종 수	-0.14	0.003
전체 백신 3차 접종 수	0.02	0.619
화이자 백신 1차 접종 수	-0.14	0.003
화이자 백신 2차 접종 수	-0.13	0.006
화이자 백신 3차 접종 수	0.03	0.531
모더나 백신 1차 접종 수	-0.08	0.099
모더나 백신 2차 접종 수	-0.08	0.077
모더나 백신 3차 접종 수	0.01	0.850
아스트라제네카 백신 1차 접종 수	-0.08	0.103
아스트라제네카 백신 2차 접종 수	-0.07	0.153
아스트라제네카 백신 3차 접종 수	0.07	0.167
얀센 백신 1차, 2차 접종 수	-0.05	0.338
얀센 백신 3차 접종 수	-0.05	0.269
노바백스 백신 1차 접종 수	0.65	0.000
노바백스 백신 2차 접종 수	0.71	0.000
노바백스 백신 3차 접종 수	0.75	0.000
기타 백신 1차 접종 수	-0.16	0.001
기타 백신 2차 접종 수	-0.18	0.000
기타 백신 3차 접종 수	-0.01	0.867
알파 변이 검출 수	-0.19	0.000
베타 변이 검출 수	-0.15	0.002
감마 변이 검출 수	-0.10	0.028
델타 변이 검출 수	-0.20	0.000
오미크론 변이 검출 수	0.64	0.000

실험 3에서 고려한 변수

변수	상관계수	유의확률
해외유입	0.02	0.784
아동기 및 청소년기 확진자 수	0.92	0.000
성년기 확진자 수	0.92	0.000
노년기 확진자 수	0.91	0.000
수도권 확진자 수	0.87	0.000
비수도권 확진자 수	0.91	0.000
사망자 수	0.80	0.000
재원 위중증 환자 수	0.80	0.000
사회적 거리두기 단계 (재구성)	0.81	0.000
코로나19 확산 관련 뉴스 기사 수	0.59	0.000
월요일 여부	-0.17	0.023
화요일 여부	0.07	0.359
수요일 여부	0.12	0.101
목요일 여부	0.03	0.646
금요일 여부	-0.01	0.899
토요일 여부	-0.03	0.668
일요일 여부	-0.01	0.847
1일 전 확진자 수	0.91	0.000
2일 전 확진자 수	0.86	0.000
3일 전 확진자 수	0.86	0.000
4일 전 확진자 수	0.85	0.000
5일 전 확진자 수	0.86	0.000
6일 전 확진자 수	0.92	0.000
7일 전 확진자 수	0.95	0.000
8일 전 확진자 수	0.87	0.000
9일 전 확진자 수	0.83	0.000
10일 전 확진자 수	0.83	0.000
11일 전 확진자 수	0.81	0.000
12일 전 확진자 수	0.80	0.000
13일 전 확진자 수	0.83	0.000
14일 전 확진자 수	0.84	0.000

변수	상관계수	유의확률
백신 1차 접종 수	0.48	0.000
백신 2차 접종 수	0.28	0.000
백신 3차 접종 수	0.75	0.000
백신 4차 접종 수	-0.09	0.253
화이자 백신 1차 접종 수	0.34	0.000
화이자 백신 2차 접종 수	-0.13	0.085
화이자 백신 3차 접종 수	0.75	0.000
화이자 백신 4차 접종 수	-0.09	0.247
모더나 백신 1차 접종 수	0.64	0.000
모더나 백신 2차 접종 수	0.52	0.000
모더나 백신 3차 접종 수	0.76	0.000
모더나 백신 4차 접종 수	-0.04	0.580
아스트라제네카 백신 2차 접종 수	0.06	0.425
아스트라제네카 백신 3차 접종 수	0.50	0.000
얀센 백신 1차, 2차 접종 수	-0.01	0.945
얀센 백신 3차 접종 수	0.70	0.000
얀센 백신 4차 접종 수	-0.02	0.821
노바백스 백신 1차 접종 수	0.77	0.000
노바백스 백신 2차 접종 수	0.59	0.000
노바백스 백신 3차 접종 수	0.28	0.000
노바백스 백신 4차 접종 수	-0.12	0.096
기타 백신 1차 접종 수	-0.13	0.079
기타 백신 2차 접종 수	-0.23	0.001
기타 백신 3차 접종 수	-0.17	0.021
중환자 전담치료병상 가동률	0.81	0.000
준-중환자 병상 가동률	0.75	0.000
감염병 전담병원 가동률	0.68	0.000
생활치료센터 가동률	0.59	0.000
누적 재택치료자 수	0.88	0.000
신규 재택치료자 수	0.87	0.000

<부록 3> 실험 1 예측 결과

실험 1에서 확진자 수와 유의한 상관관계가 있는 변수, 확진자 수와의 상관계수가 0.6, 0.7, 0.8, 0.9 이상인 변수를 차례로 고려하여 예측을 수행하면 다음과 같다. 각각의 경우에서 활용한 입력변수의 개수는 24개, 21개, 17개, 12개, 8개이다.

실험 1: 24개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	18.94	17.12	10.92	23.50	11.73	14.78	17.34	14.32	15.73
2	28.55	30.11	20.84	28.48	23.65	26.56	31.35	24.94	15.58
3	24.79	29.76	37.21	21.31	21.15	29.44	23.26	19.18	12.81
4	17.56	23.61	33.09	35.02	14.56	28.98	19.19	21.50	13.18
5	13.94	18.34	23.17	12.63	14.45	20.41	13.66	10.99	21.17
6	11.68	14.16	12.75	16.05	26.13	34.36	16.20	23.11	14.02
7	15.01	16.77	13.74	15.68	25.23	48.93	21.28	24.86	21.10
8	15.25	16.46	21.38	15.21	22.94	55.89	25.34	17.79	15.87
9	15.71	21.29	18.79	24.35	49.70	56.38	24.72	16.16	21.43
10	16.24	16.29	12.96	17.53	44.33	66.76	12.13	28.82	20.26
11	16.12	12.13	18.07	20.86	42.35	42.88	18.10	20.93	14.35
12	19.48	16.08	15.46	53.85	28.25	49.31	18.87	19.72	22.75
13	12.48	23.55	16.92	18.70	43.69	316.98	21.97	21.98	18.87
14	18.53	21.71	18.96	14.52	58.60	27.31	9.43	34.10	13.15
최소	11.68	12.13	10.92	12.63	11.73	14.78	9.43	10.99	12.81
최대	28.55	30.11	37.21	53.85	58.60	316.98	31.35	34.10	22.75
평균	17.45	19.81	19.59	22.69	30.48	58.50	19.49	21.32	17.16

실험 1: 21개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	12.68	11.36	10.86	10.70	12.66	10.22	12.83	11.89	9.58
2	13.98	12.92	13.68	12.95	15.23	11.01	12.88	12.53	11.51
3	12.07	12.79	11.96	11.65	13.37	12.81	11.03	12.63	10.40
4	11.61	12.08	12.95	14.54	12.09	19.84	13.14	12.58	11.71
5	12.74	13.94	12.58	11.04	14.43	13.94	12.53	12.14	14.06
6	14.33	11.56	12.54	13.22	13.54	13.97	12.94	14.41	15.89
7	11.24	12.20	11.75	11.63	11.73	29.48	11.65	10.75	11.86
8	11.34	10.30	11.13	10.82	12.85	30.21	11.44	11.43	10.31
9	12.27	12.31	12.28	11.22	22.00	23.94	11.27	11.73	9.65
10	12.49	12.03	11.47	9.98	14.23	41.60	11.59	11.97	14.82
11	12.86	12.37	11.62	11.37	17.16	23.85	11.03	12.14	10.26
12	11.58	11.67	11.14	10.30	14.70	56.98	13.95	10.76	13.29
13	13.51	14.34	9.72	10.67	17.76	52.34	15.00	13.74	13.88
14	12.39	11.26	11.78	10.43	13.60	33.55	13.13	14.66	11.11
최소	11.24	10.30	9.72	9.98	11.73	10.22	11.03	10.75	9.58
최대	14.33	14.34	13.68	14.54	22.00	56.98	15.00	14.66	15.89
평균	12.51	12.22	11.82	11.47	14.67	26.70	12.46	12.38	12.02

실험 1: 17개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	11.77	13.12	11.77	11.99	12.17	9.94	10.85	10.89	9.52
2	13.94	12.76	11.54	14.09	12.57	10.99	12.23	11.47	11.83
3	12.66	13.09	11.98	13.34	11.69	11.84	12.72	12.36	12.97
4	12.08	12.92	13.29	13.98	13.35	16.16	12.23	11.88	10.08
5	13.07	13.91	11.90	11.07	14.21	13.13	12.90	12.50	13.46
6	13.24	13.14	14.90	16.52	14.14	18.01	13.88	14.97	16.87
7	12.36	11.92	11.47	11.03	12.95	40.89	11.38	10.48	9.54
8	11.22	12.62	11.99	11.95	12.37	22.99	11.08	10.38	11.73
9	12.90	11.65	12.42	9.91	13.16	19.17	11.64	12.26	9.73
10	12.64	13.74	12.29	11.74	15.24	14.71	11.54	10.70	10.17
11	13.86	13.87	12.84	10.37	17.47	21.74	12.87	10.75	11.61
12	11.27	12.35	10.95	12.86	20.83	25.60	12.97	11.73	8.40
13	12.61	11.66	10.54	10.19	12.60	37.28	13.34	12.63	11.33
14	11.85	11.46	10.80	11.02	18.49	28.46	10.80	11.39	11.35
최소	11.22	11.46	10.54	9.91	11.69	9.94	10.80	10.38	8.40
최대	13.94	13.91	14.90	16.52	20.83	40.89	13.88	14.97	16.87
평균	12.53	12.73	12.05	12.15	14.37	20.78	12.17	11.74	11.33

실험 1: 12개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	12.88	12.22	11.72	11.33	11.54	10.57	12.79	11.65	11.43
2	12.35	12.22	11.54	12.94	11.55	10.35	12.60	11.22	9.51
3	12.15	12.32	10.23	12.45	10.80	9.36	12.02	12.37	11.89
4	14.17	12.87	13.86	11.55	12.13	14.46	12.12	12.91	12.53
5	13.50	14.03	11.84	11.71	11.19	13.03	12.24	14.78	9.97
6	12.89	13.38	14.02	14.46	19.55	15.20	11.74	10.69	11.26
7	13.31	12.33	12.06	9.01	11.25	17.73	10.97	10.34	10.49
8	12.93	12.97	13.18	11.22	22.39	24.20	11.52	11.53	10.69
9	13.43	12.94	12.94	12.87	16.49	32.44	12.23	11.18	10.05
10	12.98	12.35	11.39	10.43	13.30	41.05	12.54	11.81	10.83
11	14.06	11.91	12.52	9.86	12.79	25.11	10.78	9.86	11.84
12	12.84	13.06	12.78	9.53	14.44	25.74	11.84	12.42	9.97
13	13.13	14.27	11.11	10.16	18.44	46.21	13.02	10.60	11.48
14	11.96	11.88	12.41	14.97	16.77	24.42	13.12	10.34	11.66
최소	11.96	11.88	10.23	9.01	10.80	9.36	10.78	9.86	9.51
최대	14.17	14.27	14.02	14.97	22.39	46.21	13.12	14.78	12.53
평균	13.04	12.77	12.26	11.61	14.48	22.13	12.11	11.55	10.97

실험 1: 8개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	13.17	13.98	13.06	13.68	12.45	11.80	12.78	12.05	11.50
2	14.26	14.49	14.98	12.20	11.42	13.27	13.98	13.95	12.46
3	16.07	16.54	14.51	13.94	14.27	12.79	13.58	14.31	13.34
4	15.21	14.19	13.32	12.44	12.70	17.29	13.42	12.32	13.93
5	13.45	13.66	12.16	12.66	11.09	15.49	11.87	11.02	11.96
6	13.55	13.29	11.71	12.65	15.46	17.14	13.94	11.69	12.19
7	13.78	13.91	14.21	10.41	10.68	27.96	12.59	13.69	11.25
8	14.61	12.35	12.68	9.77	16.93	17.84	12.43	11.13	10.76
9	15.45	13.96	10.61	8.71	17.12	32.66	9.41	11.08	10.70
10	13.29	13.39	13.25	10.74	11.87	24.32	11.11	11.80	10.08
11	11.87	10.89	11.41	10.07	15.45	21.24	11.98	10.95	10.90
12	12.48	12.18	11.53	10.97	23.25	32.65	12.24	11.10	13.21
13	12.90	13.88	12.91	10.66	15.36	53.98	12.07	13.26	10.69
14	12.54	13.68	11.09	12.22	18.20	20.09	11.54	14.36	10.43
최소	11.87	10.89	10.61	8.71	10.68	11.80	9.41	10.95	10.08
최대	16.07	16.54	14.98	13.94	23.25	53.98	13.98	14.36	13.93
평균	13.76	13.60	12.67	11.51	14.73	22.75	12.35	12.34	11.67

<부록 4> 실험 2 예측 결과

실험 2에서 확진자 수와 유의한 상관관계가 있는 변수, 확진자 수와의 상관계수가 0.6, 0.7, 0.8, 0.9 이상인 변수를 차례로 고려하여 예측을 수행하면 다음과 같다. 각각의 경우에서 활용한 입력변수의 개수는 38개, 24개, 22개, 20개, 19개이다.

실험 2: 38개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	15.58	14.37	15.38	17.55	18.89	23.17	14.49	14.90	18.82
2	15.55	18.70	18.14	19.28	23.62	27.51	14.82	19.17	22.88
3	12.79	14.66	17.60	17.02	24.36	24.57	17.89	17.41	42.00
4	14.06	14.42	26.12	23.44	33.51	36.78	14.64	23.17	41.22
5	13.30	13.05	35.18	23.48	22.41	40.50	14.23	16.27	40.67
6	13.25	16.42	13.43	18.59	42.23	41.93	15.33	21.43	35.93
7	14.49	13.34	28.09	22.05	49.08	71.76	14.44	19.77	44.54
8	12.34	12.92	29.88	23.79	41.41	55.65	13.90	17.49	42.60
9	12.22	13.72	18.69	14.22	50.84	91.26	12.89	21.08	58.09
10	13.27	17.39	19.42	30.96	63.50	74.62	14.82	20.19	21.77
11	13.20	13.21	29.29	16.54	52.5	66.92	15.35	23.67	63.92
12	13.95	17.94	27.18	34.25	84.47	72.20	14.13	16.73	61.35
13	14.33	15.43	27.27	29.28	39.00	84.12	16.18	27.97	69.87
14	12.87	15.11	16.36	51.33	31.75	92.65	14.49	34.89	79.44
최소	12.22	12.92	13.43	14.22	18.89	23.17	12.89	14.90	18.82
최대	15.58	18.70	35.18	51.33	84.47	92.65	17.89	34.89	79.44
평균	13.66	15.05	23.00	24.41	41.26	57.40	14.83	21.01	45.94

실험 2: 24개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	11.18	11.13	11.51	11.59	12.39	12.97	11.88	11.6	11.56
2	11.48	11.13	11.77	12.06	13.72	39.59	11.04	12.67	12.00
3	11.65	11.36	11.74	11.18	11.90	14.59	11.44	10.96	12.33
4	11.90	12.19	12.08	11.59	13.02	13.56	11.41	11.48	11.85
5	12.24	12.20	11.49	11.92	13.83	18.03	11.28	11.34	11.03
6	11.60	11.79	12.53	11.72	12.02	15.57	12.05	11.51	11.10
7	12.64	12.42	12.94	11.41	12.65	16.26	11.81	10.91	12.04
8	12.66	12.05	11.45	11.12	16.54	13.58	11.30	12.19	12.32
9	11.21	12.71	12.53	11.15	14.79	30.03	11.73	11.29	11.82
10	12.86	11.67	12.61	12.99	12.83	204.49	11.66	11.87	13.55
11	11.50	12.54	11.87	13.28	16.31	41.21	11.88	11.47	13.06
12	12.84	11.91	12.64	12.27	14.98	34.56	11.58	11.87	11.66
13	11.55	14.32	10.83	11.08	13.11	25.05	10.76	11.18	13.80
14	12.05	12.95	10.94	13.08	15.34	41.17	11.45	13.79	11.26
최소	11.18	11.13	10.83	11.08	11.90	12.97	10.76	10.91	11.03
최대	12.86	14.32	12.94	13.28	16.54	204.49	12.05	13.79	13.80
평균	11.95	12.17	11.92	11.89	13.82	37.19	11.52	11.72	12.10

실험 2: 22개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	12.26	12.94	13.63	12.84	12.51	12.35	12.82	12.79	12.93
2	12.20	12.59	11.65	12.83	14.46	13.94	11.10	13.19	11.97
3	13.47	12.86	13.36	11.44	13.26	22.44	11.67	12.40	11.71
4	12.14	12.80	12.08	11.26	12.50	15.36	13.63	12.95	12.90
5	12.73	11.72	11.63	12.83	12.66	20.66	12.10	12.41	12.81
6	12.85	12.74	12.38	12.68	15.69	13.83	12.38	12.44	12.78
7	12.77	13.35	14.16	13.55	13.93	21.76	12.49	12.27	11.35
8	12.67	13.45	12.74	13.66	13.19	21.36	11.45	13.20	12.91
9	13.31	12.93	12.75	12.24	14.90	63.75	11.88	12.85	11.86
10	11.98	12.82	12.36	12.14	14.90	17.17	11.77	12.47	12.27
11	13.48	13.17	12.21	13.65	18.67	14.76	12.54	12.25	12.47
12	12.03	13.63	12.97	11.44	15.62	31.98	12.68	12.00	11.85
13	13.39	12.41	12.66	12.26	17.51	25.84	11.91	13.76	17.24
14	13.16	14.08	12.78	14.10	14.29	44.25	12.47	12.47	11.89
최소	11.98	11.72	11.63	11.26	12.50	12.35	11.10	12.00	11.35
최대	13.48	14.08	14.16	14.10	18.67	63.75	13.63	13.76	17.24
평균	12.75	12.96	12.67	12.64	14.58	24.25	12.21	12.67	12.64

실험 2: 20개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	13.91	12.57	12.15	12.27	12.46	13.03	11.97	12.16	12.18
2	12.82	13.40	11.12	13.76	13.87	13.44	12.6	14.35	12.44
3	12.00	12.60	14.22	13.16	13.23	17.49	12.78	12.71	13.08
4	11.91	12.82	12.43	12.29	13.27	19.85	11.72	11.71	13.69
5	12.52	12.06	13.47	11.65	13.21	22.07	12.68	11.84	11.91
6	13.14	13.17	13.78	13.21	12.40	19.04	13.30	11.24	11.59
7	12.76	12.71	13.94	13.82	14.74	24.61	12.00	11.64	13.53
8	12.38	14.35	12.78	12.14	14.16	41.40	11.77	11.94	12.82
9	13.14	12.68	11.97	13.08	16.45	20.41	12.53	11.94	11.78
10	12.58	12.69	12.39	11.90	15.99	42.14	11.72	12.90	14.17
11	13.61	13.94	11.61	11.88	16.64	29.09	13.40	12.57	14.06
12	12.33	12.57	11.42	12.94	17.96	26.45	12.01	13.24	13.80
13	12.76	12.78	11.12	11.85	16.21	23.40	11.69	12.16	13.70
14	11.64	13.13	12.84	13.30	13.99	23.78	11.85	12.69	11.57
최소	11.64	12.06	11.12	11.65	12.40	13.03	11.69	11.24	11.57
최대	13.91	14.35	14.22	13.82	17.96	42.14	13.40	14.35	14.17
평균	12.68	12.96	12.52	12.66	14.61	24.01	12.29	12.36	12.88

실험 2: 19개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	12.08	12.96	12.14	12.81	15.01	13.33	11.74	12.24	13.90
2	11.69	13.00	11.30	11.40	12.56	16.55	12.61	14.51	12.84
3	12.23	12.33	11.31	11.53	13.30	17.97	13.45	11.44	14.18
4	12.69	12.07	11.56	12.03	12.76	25.77	12.31	12.32	12.18
5	12.07	12.05	12.96	12.37	15.60	14.19	12.74	13.42	14.07
6	12.53	13.10	12.94	12.55	14.34	16.72	12.11	11.73	11.93
7	12.65	12.92	12.64	13.04	15.24	20.51	12.44	11.96	12.59
8	13.53	12.28	11.42	11.75	16.48	19.32	12.40	13.01	12.08
9	12.14	12.47	12.33	12.77	13.28	22.44	12.36	11.89	11.93
10	12.55	12.93	13.02	12.41	15.73	20.05	12.44	12.27	14.86
11	13.50	14.16	12.81	13.50	17.54	24.79	12.76	14.36	14.25
12	13.88	12.74	12.01	13.62	20.14	69.63	12.17	12.60	11.87
13	12.48	13.05	12.02	12.06	16.71	58.00	12.68	12.36	11.71
14	13.01	13.75	12.18	11.68	12.97	47.00	12.14	10.65	12.39
최소	11.69	12.05	11.30	11.40	12.56	13.33	11.74	10.65	11.71
최대	13.88	14.16	13.02	13.62	20.14	69.63	13.45	14.51	14.86
평균	12.64	12.84	12.19	12.39	15.12	27.59	12.45	12.48	12.91

<부록 5> 실험 3 예측 결과

실험 3에서 확진자 수와 유의한 상관관계가 있는 변수, 확진자 수와의 상관계수가 0.6, 0.7, 0.8, 0.9 이상인 변수를 차례로 고려하여 예측을 수행하면 다음과 같다. 각각의 경우에서 활용한 입력변수의 개수는 61개, 33개, 30개, 25개, 7개이다.

실험 3: 61개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	14.06	15.02	15.03	15.95	21.47	32.53	18.51	18.12	16.55
2	19.19	16.75	16.74	15.20	19.53	17.96	15.37	19.28	19.28
3	15.02	15.94	18.41	17.70	19.83	18.45	16.32	17.25	14.73
4	16.09	18.71	18.55	17.97	22.02	26.21	13.74	18.91	20.62
5	14.65	14.50	18.35	14.49	24.43	31.24	14.42	16.81	19.57
6	13.16	16.46	17.52	15.56	20.48	26.83	14.44	16.02	23.97
7	17.89	17.18	17.99	23.55	21.49	23.74	14.97	17.79	16.69
8	15.04	17.47	18.55	18.76	20.20	26.94	16.41	16.02	22.08
9	14.89	15.67	15.10	16.15	27.01	23.21	16.19	16.13	20.76
10	15.09	15.43	17.61	18.98	28.64	26.48	15.83	16.11	24.58
11	16.87	18.16	17.35	17.11	22.56	25.57	15.61	17.71	18.59
12	15.61	15.70	17.45	21.11	26.76	28.73	16.16	17.14	21.99
13	17.96	19.48	19.38	22.75	24.00	29.29	17.00	16.62	19.88
14	15.26	19.94	19.85	20.97	26.18	35.68	15.34	19.33	23.41
최소	13.16	14.50	15.03	14.49	19.53	17.96	13.74	16.02	14.73
최대	19.19	19.94	19.85	23.55	28.64	35.68	18.51	19.33	24.58
평균	15.77	16.89	17.71	18.30	23.19	26.63	15.74	17.37	20.19

실험 3: 33개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	20.46	16.59	14.87	18.22	18.46	42.13	18.22	16.44	18.00
2	18.86	18.13	15.84	20.04	17.70	26.97	18.04	20.70	18.62
3	19.01	19.88	18.58	19.54	21.73	23.37	17.54	18.63	18.11
4	17.45	19.29	18.69	17.64	19.27	29.57	18.41	17.37	21.13
5	20.06	16.62	20.28	17.65	24.19	27.53	19.25	15.99	24.71
6	18.28	20.22	21.41	17.03	19.80	27.56	17.28	15.59	19.68
7	15.53	45.59	17.36	20.33	19.58	26.93	16.83	15.75	19.58
8	17.22	28.66	20.00	20.10	23.95	23.47	17.93	17.96	19.26
9	18.68	16.46	19.58	20.76	22.97	32.11	18.69	17.68	21.77
10	18.29	19.69	19.37	18.26	28.08	29.97	19.58	17.89	22.66
11	17.34	16.15	17.34	18.89	31.11	30.20	18.32	17.08	24.21
12	18.25	18.70	21.48	23.82	24.23	29.25	18.71	18.11	20.48
13	19.32	19.67	17.78	20.69	20.76	29.82	17.34	18.83	22.09
14	20.63	22.89	20.88	17.81	26.34	44.10	19.35	16.90	23.48
최소	15.53	16.15	14.87	17.03	17.70	23.37	16.83	15.59	18.00
최대	20.63	45.59	21.48	23.82	31.11	44.10	19.58	20.70	24.71
평균	18.53	21.33	18.82	19.34	22.73	30.21	18.25	17.50	20.99

실험 3: 30개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	17.13	15.15	16.48	16.74	22.41	27.05	16.87	16.58	16.69
2	18.02	20.16	19.93	17.17	21.46	27.80	19.08	17.03	19.61
3	18.69	18.41	16.04	20.26	22.06	29.07	19.50	18.26	20.26
4	17.88	20.16	19.07	18.57	22.52	31.57	17.39	17.91	20.05
5	16.79	16.81	18.94	20.36	21.51	33.91	19.93	17.04	19.57
6	18.41	20.38	17.37	19.33	20.06	29.12	19.18	18.70	22.11
7	17.92	18.84	19.53	17.56	17.59	30.22	17.46	16.55	19.35
8	18.78	19.32	19.15	19.82	21.57	18.56	18.69	18.45	19.71
9	16.88	19.85	20.03	19.83	22.02	25.52	17.81	18.95	21.41
10	20.68	19.71	22.29	18.71	27.62	40.21	17.96	18.72	25.78
11	18.20	20.56	15.86	19.46	23.98	29.51	17.41	17.51	19.53
12	16.40	23.40	17.29	20.28	26.54	29.67	19.39	18.70	23.17
13	20.01	19.91	18.78	17.38	27.44	26.80	18.81	18.00	22.86
14	18.75	18.87	23.27	19.58	25.69	33.02	21.55	19.10	18.80
최소	16.40	15.15	15.86	16.74	17.59	18.56	16.87	16.55	16.69
최대	20.68	23.40	23.27	20.36	27.62	40.21	21.55	19.10	25.78
평균	18.18	19.39	18.86	18.93	23.03	29.43	18.65	17.96	20.63

실험 3: 25개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	24.49	16.40	15.80	13.77	20.53	27.31	17.22	16.26	15.51
2	19.15	18.10	20.85	17.96	19.30	27.31	18.27	17.98	19.76
3	17.23	17.63	26.10	19.83	21.94	24.77	19.05	19.10	17.91
4	19.47	19.39	21.40	18.15	24.42	24.08	20.21	16.37	21.12
5	18.47	17.36	17.74	19.95	22.90	33.07	19.35	16.79	20.78
6	18.21	15.98	19.27	20.16	24.36	27.50	18.32	17.66	20.77
7	18.80	18.94	19.33	19.14	24.89	24.16	20.85	15.79	18.52
8	16.96	20.96	18.67	17.00	20.12	31.89	16.79	18.28	28.62
9	18.81	21.38	21.09	20.94	29.53	29.49	20.33	17.26	21.00
10	19.75	17.34	19.25	23.65	27.67	29.14	19.44	16.29	20.20
11	18.39	18.66	19.69	31.95	25.87	43.11	20.17	18.45	23.08
12	19.07	18.03	22.47	20.90	25.19	31.18	19.06	19.27	22.56
13	20.77	19.14	24.28	17.91	25.43	28.30	19.64	18.46	21.59
14	21.49	20.82	18.86	23.42	24.17	33.83	16.51	18.52	21.98
최소	16.96	15.98	15.80	13.77	19.30	24.08	16.51	15.79	15.51
최대	24.49	21.38	26.10	31.95	29.53	43.11	20.85	19.27	28.62
평균	19.36	18.58	20.34	20.34	24.02	29.65	18.94	17.61	20.96

실험 3: 7개 변수를 반영한 모델의 예측 정확도 (MAPE, %)

Time Step	RNN 1개	RNN 2개	RNN 3개	LSTM 1개	LSTM 2개	LSTM 3개	GRU 1개	GRU 2개	GRU 3개
1	17.08	18.82	19.58	16.49	17.54	16.26	16.67	20.82	18.13
2	17.79	17.64	16.76	16.99	17.30	20.35	15.95	13.45	17.32
3	17.24	16.68	17.02	17.23	21.07	23.97	15.91	14.51	20.29
4	15.95	16.03	17.60	21.68	24.79	40.66	14.60	16.56	16.03
5	16.18	16.29	15.64	16.95	24.87	29.09	14.36	13.79	18.53
6	16.17	17.79	14.33	18.19	23.30	26.94	13.62	15.22	17.37
7	15.40	15.65	16.56	19.19	19.41	30.65	15.37	15.21	17.16
8	13.53	15.40	16.84	17.48	21.32	33.82	14.35	15.21	14.65
9	15.61	14.04	15.96	16.46	26.50	25.68	15.67	16.13	16.87
10	18.71	15.01	16.32	18.80	17.95	27.20	14.60	16.43	19.54
11	18.34	19.01	17.03	19.08	26.04	27.50	12.95	13.15	21.37
12	15.52	14.90	16.96	18.76	22.92	26.34	16.40	13.72	16.86
13	17.18	16.06	16.62	16.71	31.45	42.91	16.05	16.03	17.50
14	14.62	15.17	16.59	21.74	21.91	36.27	13.33	17.20	17.11
최소	13.53	14.04	14.33	16.46	17.30	16.26	12.95	13.15	14.65
최대	18.71	19.01	19.58	21.74	31.45	42.91	16.67	20.82	21.37
평균	16.38	16.32	16.70	18.27	22.60	29.12	14.99	15.53	17.77

ABSTRACT

Predicting Confirmed Cases of COVID-19 Using Deep Learning Models

Taekyung Hong
Department of Industrial Engineering
Sungkyunkwan University

The number of confirmed cases of COVID-19 in Korea has fluctuated rapidly. In addition, as the availability of data on COVID-19 has changed over time, it is difficult to predict it with traditional approaches such as compartment model. This paper adopted the deep learning methodology that uses past data as input variables, and utilized accessible data in each period after dividing the total analysis period. The hidden layer of the deep learning model was composed of RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory), or GRU(Gated Recurrent Unit), and the prediction of the model was evaluated by comparing MAPE(Mean Absolute Percentage Error). Specifically, the analysis period was divided considering the detection of COVID-19 mutations, and the experiments were conducted for each period: pre-mutation period, COVID-19 mutation diffusion period, and Omicron mutation diffusion period.

All experiments considered the number of confirmed cases by age group and local classification, the number of deaths and critically ill patients caused by

COVID-19, social distancing, and the number of confirmed cases in the past as input variables. Since the level of social distancing varied by time and region, this paper reconstructed it based on two common actions to prevent the spread of the virus. In each experiment, the number of news related to the spread of COVID-19 and inflows from abroad, they were used as an input variable when the correlation with the number of confirmed cases of COVID-19 was significant. In case of the number of COVID-19 mutations detected, the hospital bed occupancy rate, and the number of home caregivers, they were considered as input variables in the COVID-19 mutation diffusion experiment or Omicron mutation experiment. In each experiment, the input variables were classified according to the size of the correlation coefficient with the number of confirmed cases, and this paper confirmed that the prediction performance of the model was the best when it used certain variable.

As a result of the analysis, the best model across all experiments could not be specified as one model, but the MAPE of the best prediction in each experiment was calculated within 15%. It was confirmed that when the input variables used in the COVID-19 mutation diffusion period were applied to the Omicron mutation diffusion period, the prediction performance was worse than when the input variables were configured with the data available in this period. This study showed that the input variables of the model should be adjusted according to the change in the virus spread pattern, and suggested the need for data policy revision in order to supply consistent data stably.