https://doi.org/10.7840/kics.2023.48.4.457

스마트팜 활용을 위한 BI-LSTM 기반의 토마토 생산량 예측에 관한 연구

이 세 연•, 양 현 정*, 김 민 영**, 김 준 경***, 손 아 영°, 홍 성 훈°°

A Study on Prediction of Tomato Production Using BI-LSTM for Smart Farm Utilization

Se-Yun Lee•, Hyeonjeong Yang*, Minyoung Kim**, Junkyeong Kim***,
A-Young Son°, Seonghun Hong°°

요 약

기존 농업 분야는 최근 급격한 기후 변화와 함께 저출산 고령화로 인한 노동력 부족에 직면하고 있다. 이에 다양한 산업에 융합 및 확장되고 있는 ICT의 적용이 농업 분야에도 점차 요구되고 있다. ICT 기반의 여러 농업 기술 중 최근 확산되고 있는 스마트팜은 초기 도입 비용이 높고, 단순 생육 현황을 파악하는 정도에 그치고 있어 실사용을 위한 활용 방안을 모색할 필요가 있다. 따라서 본 논문에서는 스마트팜 인공지능 기술 도입을 위해 토마토 작물을 대상으로 생산량 예측을 위한 방법론 연구를 진행하였다. 토마토 생산량 예측 분석 모델은 수집된 데이터를 전처리하고, 변수 선택하는 과정을 거쳐 구성하였다. 여러 알고리즘 중 평가 척도를 RMSE로 했을 때 가장 효율이 좋은 BI-LSTM 알고리즘을 예측 모델에 적용하였다. 또한 스마트팜의 실사용을 위해 서비스 제공을 하고자 하였고, 웹 서비스로 구현하여 스마트팜 현장의 농부 등 사용자가 데이터를 입력 및 처리하고 해당 결과를 한 눈에 볼 수 있도록 하였다. 향후, 농업인이나 관련 연구자들은 사용자 접근성을 고려해 개선된 서비스를 제공받게 된다면, 데이터 예측 모델 뿐만 아니라 스마트팜과 관련된 다양한 연구에 널리 활용될 수 있다.

키워드 : 데이터 학습, 상관분석, 예측 분석 모델링, 스마트팜, 웹 서비스

Key Words : Data Learning, Correlation Analysis, Modeling for Prediction and Analysis, Smart Farm, Web Service

ABSTRACT

The existing agricultural sector is facing labor shortages due to low birth rates and aging population along with recent rapid climate changes. Accordingly, the application of ICT, which is converging and expanding in various fields, has become increasingly required in the agricultural field as well. Among various agricultural technologies based on ICT, smart farms that are recently spreading have high initial cost of deployment and are simply monitoring the status of growth, so it is necessary to find and prepare a plan for practical use.

※ 본 연구는 차세대융합기술연구원의 지원을 받아 수행되었습니다.(AICT-2022-0001)
• First Author : Advanced Institute of Convergence Technology, seyuncom@snu.ac.kr, 정회원
° Corresponding Author : Korea Institute of Science and Technology Information, ayson@kisti.re.kr, 정회원
°° Corresponding Author : Advanced Institute of Convergence Technology, seonghun.hong@snu.ac.kr, 정회원
* 덕성여자대학교(차세대융합기술연구원 일경험수련생), yhj990721@gmail.com
** 숭실대학교(차세대융합기술연구원 일경험수련생), min0kvv@gmail.com
*** Advanced Institute of Convergence Technology, junkyeong@snu.ac.kr, 정회원
논문번호 : 20302-019-C-RE, Received January 31, 2023; Revised March 13, 2023; Accepted March 13, 2023

Therefore, a methodological study was conducted to predict the production of tomato crops for the introduction of smart farm artificial intelligence technology. The tomato production prediction analysis model was constructed through the process of preprocessing the collected data and selecting variables. Among several algorithms, the BI-LSTM algorithm, which performs LSTM in both directions with a low RMSE value, was applied to the prediction model. In addition, it was intended to provide services for actual use of smart farms, and it was implemented as a web service so that users such as farmers on a smart farm could input and process data and see the results at a glance. In the future, farmers or related researchers will be provided with improved services under consideration of user accessibility. Then, it can be widely utilized to various studies related to not only for data prediction model but also smart-farm.

Ⅰ. 서 론

최근 빅데이터 및 인공지능 등의 기술 발전과 함께 데이터를 수집할 수 있는 센서들을 이용한 사물인터넷 (IoT, Internet of Things)의 발달로 다양한 분야에서 융합된 기술들을 활용한 산업이 확장하고 있다.

다양한 산업 분야 중에서도 농업은 코로나19의 확산 등으로 인해 식량안보 위기[1]를 맞이하였고, 우리나라에서도 농업 분야는 고령화 및 인구 감소로 인해 노동력이 감소하고 있으며, 기후 변화로 인해 농지가 유실되고 재배 산지가 북상함에 따라 농업 기반이 약화되고 있는 실정으로, 발전하고 있는 ICT 도입하여 활용하는 것이 필요하다.

이렇게 도입된 것이 스마트팜(Smart Farm) 이다. 스마트팜은 IoT 센서를 통해 농가 내/외부의 온도, 습도, 일사량, 이산화탄소, 토양 등을 측정하고 데이터를 수집할 수 있으며 이를 통해 컴퓨터, 모바일과 같은 스마트 기기를 통해 모니터링 하고, 원격 관리를 가능하게 하는 서비스를 의미한다. 나아가 농작물의 재배 과정과 유통, 품질 관리까지 하나의 플랫폼으로 관리할 수 있어 생산자와 소비자에게 정보를 제공할 수 있다.

그러나 스마트팜은 초기 비용 때문에 진입 장벽이 높은 편이며, 도입을 하더라도 단순히 센서 데이터를 확인하는 용도 위주로 사용되고 있어 실사용을 위한 활용 방안에 대한 연구가 필요하다.

우리나라 스마트팜에서 주로 키우고 있는 작물은 온실을 활용한 토마토, 딸기, 파프리카 등이다. 그 중에서도 재배 면적이 넓고, 다수의 데이터를 가지고 있는 토마토 작물을 활용하여 연구를 진행하였다.

농업 분야에서 가장 중요한 점은 생산량을 높이는 것이라고 할 수 있다. 이때, 잡초의 성장을 제거하는 것이 생산량을 높일 수 있는 가장 큰 요소[2]이지만, 본 논문에서는 농업 작물의 생산량을 예측하는 것에 초점을 맞추었다. 생산량을 예측할 수 있으면, 수확에 필요한 노동력과 포장 및 운반 등에 필요한 재료와 운송 수단을 미리 선점할 수 있어 농업 생산성에 효율적이라고 할 수 있다.

이로인해 스마트팜 연구들 중 생육 데이터를 분석하는 연구들이 진행되고 있었다. 스마트팜 데이터를 이용하여 토마토 최적인자에 관한 연구[3]를 진행하였고, 다중 회귀 분석을 통해 수확량과 생육데이터의 연관성을 연구[4]하였다. 또한 머신러닝과 딥러닝 접근법을 비교하여 토마토 수확시간을 예측[5]하기도 하였다.

연구를 위해 토마토 작물의 생육 데이터들을 수집하여 생산량 예측 모델을 구성하고자 하였다. 생산량을 예측함에 필요한 데이터들을 수치화하고 모델을 개발한다면, 추후 의사결정 지원 모델을 만드는데 도움이 될 것으로 예상된다.

또한 토마토 생산량 예측 모델을 만들고, 실험한 것에 그치지 않고 실사용을 할 수 있는 웹 기반 형태의 서비스를 구현하였으며 추후 일반 사용자에게 확대 적용하여 스마트팜 실사용을 쉽고 편리하게 할 수 있도록 할 예정이다.

본 논문에서는 토마토 생산량 예측 모델 연구와 함께 스마트팜의 실사용을 위해 다음과 같은 목표를 갖는다. 첫째, 스마트팜 수확 작물(토마토) 생산량 예측을 위한 방법론 연구를 진행하고 둘째, 스마트팜 서비스 실사용을 위해 구현한 토마토 생산량 예측 웹 서비스를 제공하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장 관련 연구 및 방법론에서 스마트팜 작물에 대한 기존 연구와 생산량 예측을 위한 방법론에 대해 기술한다. 3장 본론에서는 토마토 생산량 예측을 위한 연구 방법 및 과정을 기술하고, 실제 구현한 웹 기반의 서비스를 보인다. 4장 실험 및 평가에서 토마토 생산량 예측 실험을 진행한다. 5장 결론에서 스마트팜 데이터 분석 활용 방안과 향후 연구 계획에 대해 제시한다.

Ⅱ. 관련 연구

2.1 기존 스마트팜 관련 연구 분석

전 절에서 살펴본 바와 같이 스마트팜 관련 연구는 다양한 분야에서 이루어지고 있으며, 최근에는 생산량을 높이기 위해 잡초 제거 동작을 수행하는 로봇과 융합된 연구[6]도 진행 중임을 알 수 있었다.

기존 스마트팜 관련 연구 분석을 위해 국내 최대 학술 데이터베이스[7]를 통해 “스마트팜” 키워드로 검색한 결과, 최근 1년 동안의 국내 논문 기준으로 264건이 있다. 대부분은 현황, 방안[8,9]과 같은 현재 스마트팜 시장이나 기술력에 대해 분석한 논문들이 30건으로 11.36%를 차지하고 있었으며 스마트팜 플랫폼, 시스템 설계[10,11]에 관한 논문이 19건, 7.19%로 그 다음을 이루었다. 마지막으로 생장을 예측[3-5,14-16]하거나 질병을 탐지[12,13]하는 등 딥러닝을 활용한 논문들도 16건, 6%로 다수 있었다.

앞서 살펴본 스마트팜 관련 연구들 중, 본 논문과 연관성이 있는 것으로 판단되는 작물의 생육과 관련된 연관성 분석이나, 생산량 예측을 하기 위한 연구들을 표 1에 분석하였다.

표 1. 관련 연구 분석
Table 1. Related research ananlysis

	Purpose	Advantages	Limitations
[3]	Finding optimal factors for smart farm data	Finding the optimal factor for tomato production	Only one farmhouse was analyzed, and growth variables were not considered
[4]	Correlation between smart farm data and production	Analyze the correlation between smart farm data and production through multiple regression analysis	Doesn’t consider environment variables
[14]	Correlation between smart farm data and production	Analyze the correlation between smart farm data and production	Only one farm was analyzed, the data collection period was unclear, and environmental variables were not considered
[15]	Production and growth prediction	Predict production and growth using convLSTM	A small number of farms were analyzed, and variable selection was not performed

표 1을 분석한 결과 [3], [14]과 같이 농가 하나만을 분석하거나, [15]와 같이 소수의 농가를 분석하게 된 것은 스마트팜 데이터가 유의미하게 수집되고 있지 못함을 의미한다. 또한 [3]은 생육 변수를 고려하지 않았고, [4], [14]은 환경 변수를 고려하지 않고 데이터 분석을 수행함으로써 연구에 한계가 있었다.

본 논문에서는 관련 연구들을 분석하여 더 높은 정확도를 위해 최대한 많은 농가의 데이터를 사용하고나 하였고, 환경 변수와 생육 변수 모두를 고려한 토마토 생산량 예측 모델을 만들고자 하였다.

2.2 토마토 생산량 예측을 위한 방법론

토마토 생산량 예측 분석을 위해 수집되는 데이터는 특정 시간에 측정된 데이터가 대부분으로, 작물의 성장 흐름 및 계절의 변화에 따라 수집된 데이터이다. 이를 고려하여 다변량 시계열 예측에 활용되고 있는 알고리즘을 사용[16]하였다.

대표적인 것이 RNN(Recurrent Neural Network)을 기반으로 1997년에 Hochreiter, et al.가 제안한 LSTM (Long-Short Term Memory)이 있다. 그러나 은닉 계층에 과거의 데이터 정보를 기억하고 있어 시계열 예측에 적합하지만, 입력을 시간 순서대로 하여 결과물이 직전 패턴을 기반으로 수렴한다는 한계[17]가 있다.

이런 단점을 해결하기 위해 나온 알고리즘이 양방향을 이용하는 RNN, BI-LSTM(BI-Long Short Term Memory)[18]이다.

또한 LSTM보다는 단순한 형태로 2014년에 Cho, et al.에 의해 고안된 GRU(Gated Recurrent Unit) 알고리즘은 데이터 양이 많지 않은 스마트팜 분석에 더 적합할 것으로 예상되어 실험에 활용하였다.

4장 실험 및 평가에서 방법론에 제시된 LSTM, BI-LSTM, GRU 이 3가지 알고리즘을 통해 성능을 비교하고, 가장 적합한 알고리즘을 선정하여 토마토 생산량 예측 모델을 구성하였다.

2.3 BI-LSTM

본 논문에서는 데이터 분석 알고리즘들을 실험 및 평가를 거쳐 최종적으로 토마토 생산량 예측에 적합한 BI-LSTM 알고리즘을 선정하여 적용하였다.

다음 그림 1은 BI-LSTM의 구조이다. 기존에 많이 사용하던 RNN이나 LSTM은 결과물이 직전 패턴을 기반으로 하여 수렴하는 한계가 있어 이를 해결하는

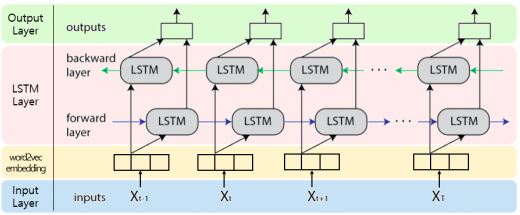


그림 1. BI-LSTM 알고리즘의 구조
Fig. 1. Structure of BI-LSTM algorithm

방법으로 양방향 LSTM(Bidirection-LSTM) 알고리즘이 제안되었다. BI-LSTM은 정보를 역방향으로 전달하는 은닉 계층(Hidden Layer)을 추가하여 정보를 보다 유연하게 처리한다[17]는 특징이 있다. 이는 각 시점에서 은닉 상태가 이전 시점과 미래 시점의 정보를 모두 갖는 효과가 있으며 각 시간 간격에서 전체 시계열로부터 학습할 수 있도록 하는 경우 유용할 것으로 기대되어 토마토 생산량 예측 알고리즘에 활용하였다.

Ⅲ. 본 론

본 장에서는 토마토 생산량 예측을 위한 연구 과정을 기술한다. 1절에서는 데이터 수집 및 전처리 과정과 분석 흐름도를 나타내고, 2절에서 변수 선택을 위한 과정과 회귀 모델들의 비교를 통해 회귀식을 도출한다. 3절에서는 스마트팜 서비스를 위해 토마토 생산량 예측 서비스 제공을 위한 웹 페이지를 보이며 마친다.

3.1 토마토 생산량 예측 모델 설계

3.1.1 토마토 생산량 예측 분석 흐름도

데이터 분석을 위해서는 설계, 준비, 가공, 분석, 결론 도출의 과정을 거친다. 이때 토마토 생산량 예측 분석을 위해서는 크게 수집된 데이터를 전처리하고, 변수를 선택하고 회귀식을 도출하여 적합한 생산량 예측 모델을 만들고, 이를 평가하는 과정이 필요하다.

다음 그림 2에서 토마토 생산량 예측 모델 분석 흐름을 보인다.

데이터 전처리(Data Preprocessing) 과정에서는 보간법, 이상치 판별 및 제거, 스케일링을 수행하였고, 많은 데이터를 확보하고 분석에 알맞은 방향으로 설정하고자 하였다.

변수 선택(Variable Selection) 과정에서는 다중공선성을 제거하고, 단계별 선택법을 통해 변수를 선택하고, 다중 회귀식을 도출하였다.

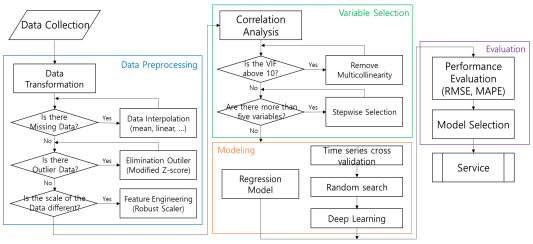


그림 2. 토마토 생산량 예측 모델 분석 흐름도
Fig. 2. Flow chart of tomato production prediction model analysis

데이터 모델링(Modeling) 과정에서는 앞서 살펴본 방법론에 따른 알고리즘을 사용하였고, 평가(Evaluation)를 위해 평가 지표를 설정하여 분석하였고, 최종 모델을 선정하여 서비스로 제공할 수 있도록 하였다.

3.1.2 토마토 데이터 수집 및 정제

본 논문에서 활용된 토마토 관련 스마트팜 데이터는 공공 데이터 포털에서 제공하는 "농촌진흥청_스마트팜 우수농가 공개용 데이터"[19]에서 완숙 토마토 관련 데이터를 사용하였다.

환경 정보로는 내부온도, 외부온도, 내부습도, 내부 CO_2, 풍속, 누적일사량, 급액 EC, 급액 pH, 일 급액횟수, 일 급액량, 1회 급액량이 제공된다. 생육 정보로는 생장길이, 초장, 엽수, 엽장, 엽폭, 엽병장, 관부직경, 줄기굵기, 개화화방, 꽃개수, 착과수, 수확수, 열매수 등의 정보가 제공되고 있다.

본 연구에서는 스마트팜의 내부 환경에 주목하여 독립 변수 중 외부 환경 요소(외부 온도)는 제외하였고, 데이터가 제대로 수집되어 있지 않은 변수(개화마디, 착과마디, 수확마디, 관부직경, 초장, 엽병장, 풍속, 급액 EC, 급액 pH, 일 급액횟수, 일 급액량, 1회 급액량)를 제외하였다. 데이터 분석에서의 종속변수는 열매 수로 설정하였다. 이후 토마토 데이터 1년치 전체를 기준으로 하여 주간 단위로 통합하였다.

공공 데이터 포털을 통해 수집하여 데이터가 어느 정도 갖춰진 농가는 50개였으며, 이를 정제하기 위해 농가를 선택하고, 결측치 및 이상치 처리를 진행하고, 스케일링을 통해 학습 모델을 위한 데이터를 정제하였다.

농가 선택 기준으로는 첫째, 개화군 데이터가 시작하는 시점을 기준으로 26~28주의 데이터를 가지고 있는 농가를 선택하였다. 토마토는 환경 조건이 좋을 경우 개화 후 40~50일 경이면 수확기에 달하고 환경이 불량하면 과실비대 및 수확까지 개화 후 70~90일

(약 10~12주)이 소요[3]된다. 또한 토마토의 수확기는 3월에서 6월까지 약 4개월(약 16주)이다. 최종적으로 개화 시작부터 수확기가 끝나는 시점인 26~28주를 농가 선택의 기준으로 삼았다. 둘째, 결측치가 연속 3주 이상 존재하는 다량의 손실데이터가 포함되어있는 농가는 제외하였다. 해당 기준을 바탕으로 13개의 농가를 선정하여 데이터셋을 구성하였다.

데이터셋에서 환경 변수 중 누적일사량의 결측치는 해당 농가와 지역과 날짜가 같은 농가 데이터로 대체하였고, 생육 변수 결측치는 선형보간법으로 대체하였다.

이상치는 모델 학습에 큰 영향을 미치고 모델 성능을 저하시키기에, 토마토 데이터를 전처리할 때에 이상치의 영향을 최대한 줄이고자 하였다. 이상치 탐색 방법으로는 수정된 표준화 점수(Modified Z-score)[20]를 사용하였으며 이를 구하는 수식은 (1)과 같다.

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \quad (1)$$

이상치의 영향을 크게 받는 평균과 표준편차를 사용하는 표준화 점수와 달리, 수정된 표준화 점수는 중앙값과 중앙값 절대편차를 이용해 이상치에 강건하다는 특성이 있다. 본 논문에서는 수정된 표준화 점수의 절댓값이 4.5보다 큰 경우를 이상치로 판단하고, 해당 이상치를 상한값 또는 하한값으로 대체하였다.

앞서 구성한 데이터셋에서 이상치 개수가 가장 많았던 한 농가를 제거해 12개 농가의 데이터를 가지고 데이터셋을 구성하였다.

마지막으로 각 변수들의 단위가 다르기 때문에 스케일링을 통해 모든 변수들의 중요도를 동일하게 설정하였다. 스케일링 방법으로는 중앙값을 사용해 이상치에 강건한 Robust Scaler[21]를 사용하였다.

3.2 토마토 데이터 변수 선택 및 회귀식

3.2.1 토마토 데이터 변수 선택 과정

토마토 생산량 예측의 정확도를 높이기 위해 변수 선택[22] 과정을 거쳤다.

반응 변수와 관계가 없는 변수들을 설명 변수로 사용하면 모델의 성능이 악화될 수 있기 때문에 피어슨 상관계수를 통해 변수 간의 상관관계를 확인해보았다. 일반적으로 상관계수의 절댓값이 0.2보다 크면 특성 간에 상관관계가 있다고 판단한다. 따라서 본 연구에서는 반응 변수와 설명 변수 간의 상관계수 절댓값이 0.2을 넘는 변수를 선택하였다.

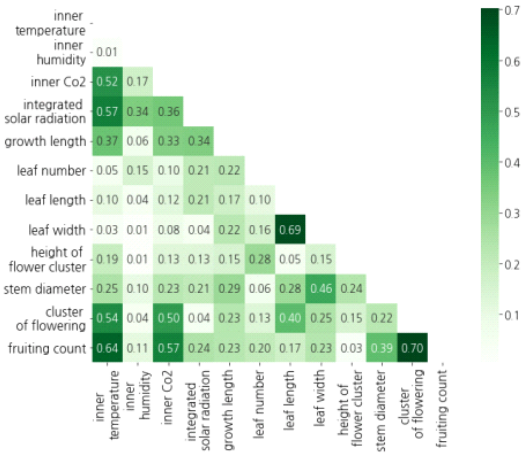


그림 3. 토마토 데이터 변수 간 상관 관계 분석
Fig. 3. Correlation analysis between tomato data variables

토마토 열매 수와 환경 변수, 생육 변수 간의 연관성을 분석하기 위해 상관분석을 수행해본 결과는 그림 3과 같다.

상관 관계 분석 결과 내부온도, 내부CO2, 누적일사량, 생장길이, 엽수, 엽폭, 줄기굵기, 개화화방이 열매수와 상관관계가 있다고 할 수 있다.

상관 관계 분석 후, 회귀 분석을 하기 위해서는 설명 변수들이 독립적이어야 한다. 설명 변수 간의 상관관계가 강하게 나타나는 경우 다중공선성이 의심되므로 일반적으로 분산팽창요인(VIF)을 계산하여 10이 넘는 경우 문제가 있다고 판단한다. VIF 결과는 표 2와 같으며, 각 변수들 모두 10 이하로 다중공선성이 나타나지 않은 것으로 나타났다.

다중공선성 이후, 의미 있는 변수를 가려내기 위해 변수 선택법을 사용하였다.

변수 선택법 중 기존 변수의 중요도가 낮아지면 변

표 2. 변수들의 분산팽창요인(VIF)
Table 2. Variance Inflation Factor(VIF) of Variables

Variables	VIF Factor
Innder Temperature	2.55
Cluster of Flowering	1.90
Inner CO_2	1.70
Leaf Width	1.44
Stem diameter	1.43
Growth Length	1.40
Number of Leaf	1.19

수를 제거하여 변수를 선택하게 되는 단계적 선택법[23]을 사용하여 회귀 분석 모형에 포함하였다. 단계적 선택법에서 Adj R Squared(=0.65)가 더 이상 높아지지 않는 구간의 변수를 채택하였고, step 6에서 상수항을 제외한 개화화방, 내부온도, 줄기굵기, 엽수, 내부CO2를 최종 설명 변수로 선정하였다.

3.2.2 회귀 모델 및 회귀식

본 논문의 목표 값인 종속 변수(열매 수)는 실수형 데이터이므로 값 예측을 위해 회귀 모델을 사용하였다.

회귀 모델은 한 개 이상의 독립 변수(X)와 종속 변수(y)와의 선형 상관 관계를 모델링하는 회귀분석 알고리즘인 다중 선형 회귀(Linear Regression)를 사용하였다. 또한 특정 조건에 따라 분기를 만들어 가지를 뻗어나가고, 그 결과를 만족하는 최종 노드의 독립 변수 평균값으로 값을 반환하는 모델인 의사 결정 나무 기반의 회귀 알고리즘 의사 결정 나무(Decision Tree Regression), 랜덤포레스트(RandomForest), Tree를 기반으로 하면서 속도가 빠른 편인 LGBM(Light Gradient Boosting Model), 예측력이 좋아서 많이 사용되고 있는 XGB(Extreme Gradient Boosting Regression)을 각각 사용하여 총 다섯 가지의 회귀 모델의 평가 지표를 비교하였다.

평가 지표는 모델 신뢰성 검증을 위해 직관적으로 볼 수 있으며 에러의 왜곡이 줄어드는 평균 제곱근 오차(RMSE, Root Mean Square Error)와 오차 평균의 크기가 더 작은 모델을 좋은 모델로 평가하는 평균 절대 비율 오차(MAPE, Mean Absolute Percentage Error)를 사용하였으며, 비교는 다음 표 3과 같다.

비교 결과 다중 선형 회귀 모델의 평가 지표가 가장 낮아 정확도가 가장 높다고 볼 수 있었다. 이를 활용하여 회귀식을 식 (2)과 같이 나타내었다.

$$y = -0.149 + 0.728x_1 + 0.253x_2 + 0.210x_3 - 0.346x_4 - 0.494x_5 \quad (2)$$

식 (2)에서 y는 토마토의 수확량이며, x_1은 개화화방, x_2는 내부CO2, x_3 엽수, x_4 줄기굵기, x_5는 내부온도를 나타낸다. 이를 통해 토마토 수확량은 개화화방, 내부CO2, 엽수와는 양의 상관관계를 가지고, 줄기굵기, 내부온도와는 음의 상관관계를 가지고 있음을 알 수 있었으며 토마토 수확량에 이와 같은 변수들이 영향을 줄 수 있음을 알아내었다.

표 3. 회귀 모델의 평가 지표 비교
Table 3. Comparison of evaluation metrics for regression models

Regression models	Evaluation Metrics	Value
Linear Regression	RMSE	1.952
	MAPE	0.082
Decision Tree Regression	RMSE	2.282
	MAPE	0.105
RandomForest	RMSE	2.098
	MAPE	0.091
LGBM	RMSE	2.379
	MAPE	0.106
XGBoost	RMSE	1.963
	MAPE	0.084

3.3 토마토 생산량 예측 서비스를 위한 웹 구현

토마토 생산량 예측 서비스를 다수의 사람들이 쉽게 접근하여 사용할 수 있도록 서비스 형태로 제공하고자 하였다. 컴퓨터/스마트 기기간 호환을 고려하여 웹 페이지 형태로 구현하고자 하였다.

웹 개발 프레임워크는 데이터 분석 모델에 사용한 Python을 기반으로 웹 개발에 많이 사용하고 있는 Django를 활용하였다. 다음 그림 4는 Django가 동작하는 기본 구조를 나타낸다.

사용자는 스마트 기기나 컴퓨터를 통해 웹 서비스에 접속하고, Django를 거쳐 서비스를 이용할 수 있게 된다. View는 데이터를 가져오고, 결과 값을 전달하는 역할을 하며 Template은 실제 사용자에게 보여지는 화면을 구성하고 있으며, Model은 데이터베이스와 연결하는 역할을 한다.

웹 페이지는 별도의 가입 절차 없이 이용할 수 있도록 네이버 로그인을 통해 인증을 할 수 있도록 하였고, 토마토 농가의 데이터를 입력하면 그 값을 알아보기 쉽게 표 형태로 시각화하여 구성하여 확인할 수 있

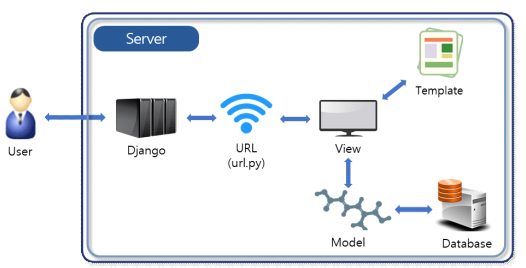


그림 4. Django의 기본 구조도
Fig. 4. Django basic structure

도록 하였다.

사용자로부터 입력받은 데이터와 분석을 수행한 결과 데이터는 지속적인 학습을 위한 포맷으로 저장하여 누적될 수 있도록 하였다.

또한 사용자 편의를 위해 이미 구축되어 있는 스마트팜 내 저장되는 센서 데이터들은 그대로 웹 서비스에 사용할 수 있도록 파일을 통한 업로드가 가능하도록 확장성을 고려하였다.

다음 그림 5는 토마토 생산량 예측 서비스 이용을 위한 웹 페이지의 화면이다.

웹 페이지에서 설명 변수로 선정된 내부 온도, 내부 CO_2 및 개화화방, 줄기굵기, 엽수 데이터를 입력 받고 입력한 데이터를 확인할 수 있도록 나타내었다.

토마토 생산량 예측 분석 결과 화면은 다음 그림 6과 같다. 입력된 데이터와 기존에 있던 데이터를 통해 분석을 수행한다.

결과 페이지를 통해 토마토 생산량 예측 개수를 보여주고, 스마트팜 내에서 조절이 가능한 내부 온도와 내부 CO_2 값을 우수 농가의 상한/하한 값을 비교군 그래프로 제공하여 사용자 농가 환경과 비교할 수 있도록 하였다. 이때 우수 농가의 상한/하한 값은 공공

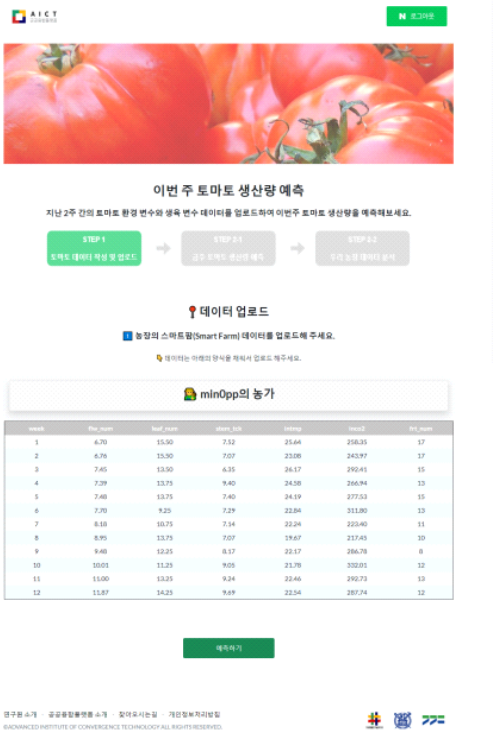


그림 5. 토마토 생산량 예측 웹 서비스 화면 I
Fig. 5. Tomato production prediction Web service screen I

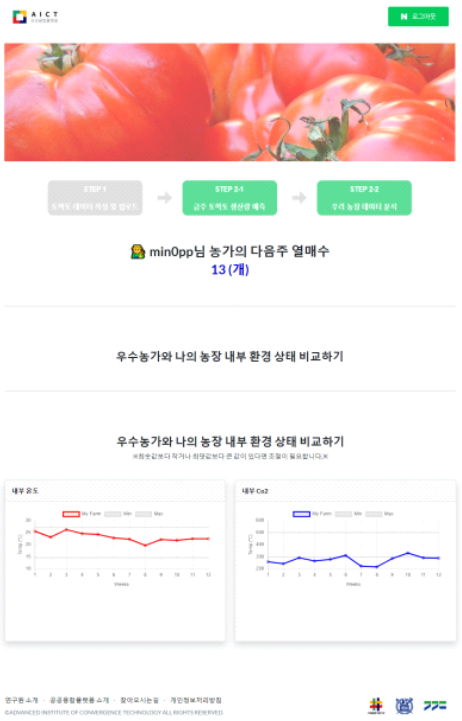


그림 6. 토마토 생산량 예측 웹 서비스 화면 II
Fig. 6. Tomato production prediction Web service screen II

데이터 우수 농가[19]의 완숙 토마토 부분 데이터 중, 수확량이 많은 농가의 평균을 구하여 활용하였다.

웹 페이지를 통해 농가는 우수 농가의 데이터를 가이드라인으로 활용할 수 있으며, 농가 예측 값과 비교하여 열매 예측 값이 낮을 경우 제시된 환경 상태에 따라 농가 내부의 온도, CO_2 값을 조절하여 생산성 향상을 기대할 수 있다.

Ⅳ. 실험 및 평가

4장에서는 토마토 생산량 예측 분석 모델에 대한 실험 및 평가를 진행한다. 1절에서 실험을 진행한 환경 및 구성에 대해 설명하고, 2절에서 토마토 생산량 예측 모델 성능 평가를 수행하여 BI-LSTM 알고리즘의 성능을 보인다.

4.1 실험 환경

실험은 앞서 3장에서 구성한 데이터셋을 활용하였다.

토마토를 심은 후의 주간 값을 데이터 분석을 위한 인덱스 값으로 지정하였고, 변수 선택을 통해 선정된

환경 변수 2개(내부온도, 내부 CO_2)와 생육 변수 3개(개화화방, 줄기굵기, 엽수)를 사용하였다.

다음 표 4는 토마토 생산량 예측을 위해 분석 학습에 활용한 실험 환경으로 Python 활용을 위해 물리적인 컴퓨터 사양과 가상 머신 환경으로 사용한 환경을 나타낸다.

표 4. 실험 환경
Table 4. Evaluation Environment

Environment		Specifications
Physical Environment	CPU	AMD Ryzen 5 5600X 6 Core Processor 3.70 GHz
	RAM	16GB
	OS	Windows 10 (64 bit)
IDE (Integrated Development Environment)	Interface	Jupyter Notebook
	Notebook Server	Version 6.0.2
	Python	Version 3.7.10
	TensorFlow	Version 2.1.0

4.2 토마토 생산량 예측 모델

본 연구에서는 토마토 생산량 예측을 위해 3가지 인공신경망 알고리즘을 사용해 가장 성능이 좋은 모델을 알아보고자 하였다.

토마토 생산량 예측을 위해, 앞서 방법론에서 살펴본 바와 같이 LSTM, BI-LSTM, GRU를 사용하였으며 각 알고리즘별 토마토 생산량 예측을 위한 데이터 분석 모델 흐름도는 다음 그림 7과 같다.

데이터셋은 학습을 위해 트레이닝 데이터셋 80%, 테스트 데이터셋 20%로 나눠 진행하였으며, 트레이닝 데이터셋에서 20%는 검증을 위해 구분해두었다.

데이터셋을 나눈 후, 예측 값을 정확하게 하기 위한 하이퍼파라미터 튜닝을 진행하였다. 가능한 조합을 시도하여 최적의 하이퍼파라미터 값을 튜닝하는 Random Search[24]를 사용하여 하이퍼파라미터 값을 정교하게 설정하고자 하였다.

튜닝 시 데이터 분석 모델의 신뢰도를 높이고자 교차 검증을 진행하고자 하였으며, 시계열 데이터 모델링에서 교차 검증을 위해 많이 사용하는 TimeSeriesSplit[25]을 사용하였다.

농가마다 Random Search를 실행해 가장 좋은 test score 값이 나온 파라미터를 파악한 뒤, 12개 농가에서 공통적으로 나온 파라미터 조합들을 파악하였다. 해당 조합들을 대상으로 모델링하여 12개 농가에서의

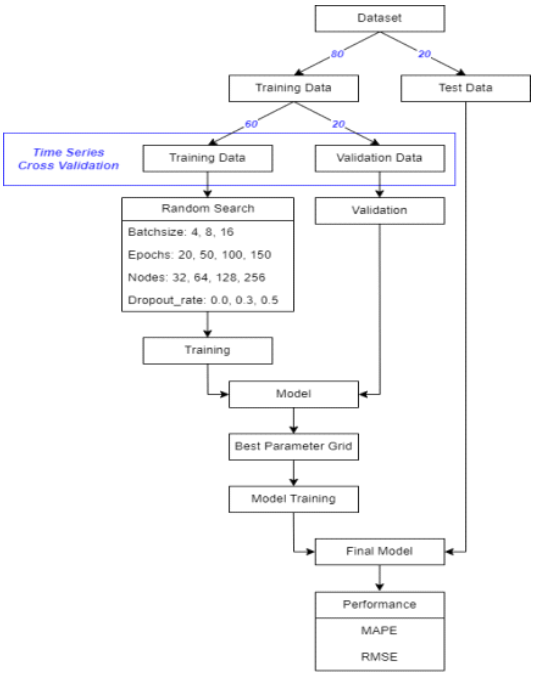


그림 7. 토마토 생산량 예측을 위한 데이터 분석 모델 흐름도
Fig. 7. Flow Chart of data analysis model for tomato production prediction

평가 지표 값의 평균이 가장 좋은 조합으로 최종 하이퍼 파라미터를 설정하고자 하였다.

활성화(Activation) 함수는 relu, 손실(loss) 함수는 mse, 최적화(Optimizer) 함수는 adam으로 설정하였다. Batchsize는 4, 8, 16으로, Epochs은 20, 50, 100, 150으로, Nodes는 32, 64, 128, 256으로 Dropout_rate는 0, 0.3, 0.5로 설정하여 각각 실험을 진행하였다.

모델 성능 평가 지표로는 RMSE와 MAPE를 사용하였다. RMSE의 경우 예측값의 크기에 의존적이므로[26] 다양한 규모의 농가에 대한 사용성을 높이고자 백분율 표현의 MAPE를 함께 사용[27]하였다. 이때 RMSE와 MAPE는 값이 낮을수록 예측 값이 정확하다는 것을 의미한다.

모든 평가 지표는 테스트 데이터셋으로 산출하여 12개의 농가의 값에 대한 평균으로 성능을 평가하였고 실험 결과는 표 5와 같다.

BI-LSTM은 하이퍼파라미터 튜닝시 Node가 256, Epochs을 50, Dropout은 0.5, Batchsize는 16으로 했을 때 가장 좋은 성능이 나타남을 알 수 있었으며, 표 5에 각 알고리즘별 최적의 하이퍼파라미터 튜닝 값을

표 5. 열매수 예측 모델 실험 결과
Table 5. Result of fruiting count prediction model

Models	Hyperparameters	Evaluation Metrics	Value
BI-LSTM	node:256, epoch:50, drop:0.5, batch:16	RMSE	1.952
		MAPE	0.082
GRU	node:64, epoch:50, drop:0.0, batch:16	RMSE	2.282
		MAPE	0.105
LSTM	node:32, epoch:150, drop:0.0, batch:16	RMSE	2.098
		MAPE	0.091

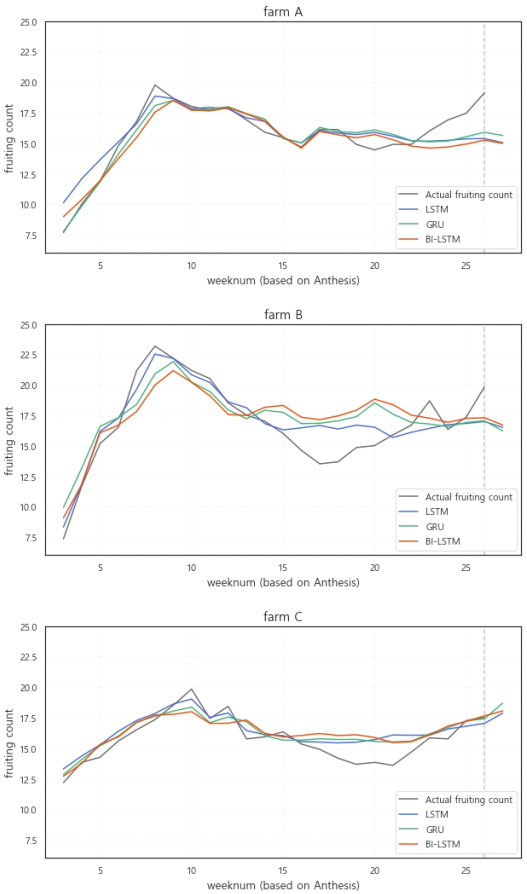


그림 8. A, B, C 농가의 알고리즘별 생산량 예측 그래프
Fig. 8. Production Prediction graph by algorithm

나타낸다.

실험 결과 BI-LSTM 수행시 RMSE가 1.952, MAPE가 0.082로 낮게 나타나 다른 알고리즘보다 성능이 좋음을 보였다.

최종적으로 토마토 생산량 분석 모델에 BI-LSTM 알고리즘을 통해 서비스로 제공하는 웹에서 토마토 생산량 예측 결과 값을 제공한다.

다음 그림 8은 3개 농가의 알고리즘별 토마토 생산량 예측 그래프를 나타낸다.

제공하는 서비스는 2주 동안의 데이터를 학습해 다음 1주를 예측하는 모델이므로, 주차를 나타내는 X축은 첫 예측 주차인 3부터 시작한다. 검정색은 실제 농가에서 측정한 토마토 생산량이고, 파란색은 LSTM, 초록색은 GRU, 빨간색은 BI-LSTM 알고리즘을 활용하여 예측한 토마토 생산량 값을 나타낸다.

실제 농가 측정 값과 많이 차이가 나는 주차는 꽃이나 잎을 임의적으로 제거하는 등 데이터 학습으로 예측하지 못하는 경우인 것으로 판단된다. 차후 생산량이 감소하는 부분에 대한 데이터 학습을 통해 오차율을 줄여나가고자 한다.

웹 서비스를 통해 간편하게 이용할 수 있도록 최소한의 데이터 입력을 받고자 2주 간의 데이터 입력을 하여 분석을 진행할 수 있도록 하였다.

2주 동안의 데이터를 통해 다음 1주의 토마토 생산량 예측 값을 구하도록 하였고, 입력한 데이터들을 데이터베이스에 저장하였다.

서비스를 제공하며 데이터가 쌓여 학습을 진행할수록 예측 값은 정확해질 것이며 이에 따라 사용자들의 신뢰도가 높아질 것이다.

이렇게 예측된 토마토 생산량을 통해 농가는 실내 온도나 습도, CO2를 조절해 생산량을 높이거나, 잎을 솎아내거나 꽃을 따주는 등의 작업을 하여 이후 수확량에 대한 조절이 가능할 것으로 기대된다.

V. 결론 및 향후 과제

최근 식량안보 위기, 노동력 감소, 기후 변화 등으로 인한 농업의 문제점과 ICT, IoT 기술 발전을 통해 스마트팜 도입이 필요해짐에 따라 농업 관련 기술과 데이터 분석의 중요성이 커지고 있다.

이미 스마트팜 관련 연구들이 활발하게 진행됨에 따라 다양한 방법론과 제품들이 제시되고 있지만, 이를 실사용하기 위해서는 먼저 접근이 쉬워야하고, 신뢰할 수 있는 결과 값을 도출하는 것이 필요하다.

본 논문은 토마토 생산량 예측 모델 연구 및 스마트팜 서비스의 실사용을 위해 다음과 같은 목표를 가지고 수행하였다.

첫번째, 스마트팜 수확 작물(토마토) 생산량 예측을 위한 방법론 연구 : 토마토 생산량 예측을 위해 데이터 분석 모델을 구성하였다. 기존의 연구들과는 달리 환경 변수와 생육 변수를 모두 고려하였고, 다수의 농

가 특성을 고려하여 수집한 데이터들을 최대한 반영하고자 하였다. 정확도 높은 결과를 도출하게 위해 전처리, 변수선택 과정을 수행하였고, 회귀식을 도출하는 과정을 거쳤다. 또한 최적의 하이퍼파라미터 값으로 튜닝하고자 가능한 조합을 시도해보는 Random Search를 사용하였고, 교차 검증을 통해 신뢰도를 높이고자 하였다.

이를 통해 스마트팜 관련 데이터 분석 모델 뿐만 아니라 다양한 데이터 분석 모델에 대한 방법론으로 활용될 수 있을 것으로 기대된다.

두번째, 스마트팜 서비스 실사용을 위한 웹 기반의 토마토 생산량 예측 서비스 제공 : 스마트팜 연구가 실제 농가에 적용되어 사용하기 위해서는 접근이 쉬워야한다고 판단되어, 웹 기반 서비스 제공을 목표로 하였다. 입력을 최대한 쉽고 편리하게 할 수 있도록 구성하였으며, 토마토 생산량 예측 분석 결과 값이 사용자에게 쉽게 보일 수 있도록 시각화하여 제공하는데 초점을 두었다. 실험 및 성능 평가를 통해 BI-LSTM 알고리즘을 적용하였고, 토마토 생산량 예측에 적합한 알고리즘을 제공함으로써 정확도 높은 서비스를 제공할 수 있도록 하였다.

향후 지속적인 학습을 통하여 토마토 생산량 예측 모델의 오차를 줄여 정확도를 높일 수 있도록 데이터를 수집 및 분석하는 작업을 계속 해나갈 것이며, 다른 작물에도 도입하여 실증 연구를 진행할 계획이다. 또한 구현한 웹 서비스를 보다 많은 사람이 접근하여 사용할 수 있도록 개선하고자 한다.

References

[1] Y. Jang, S. Son, and K. Ryou, "Food security policies of GCC countries and their implications for Korea in the Post-COVID-19 Era," *Studies in Global and Regional Strategies*, vol. 20, no. 8, Dec. 2020. (ISBN : 978-89-322-9023-2 94320)

[2] N. Jamil, G. Kootstra, D. F. van Apeldoorn, E. J. Van Henten, and L. Kooistra, "Investigating the effect of strip treatment and border-row effect on cabbage growth in intercropping system using time series of UAV imagery," *The XX CIGR World Congress 2022*, Dec. 2022.

[3] M. H. Na, Y. Park, and W. H. Cho, "A study on optimal environmental factors of tomato using smart farm data," *J. Korean Data & Inf. Sci. Soc.*, vol. 28, no. 6, pp. 1427-1435, 2017. (https://doi.org/10.7465/jkdi.2017.28.6.1427)

[4] A. Hong, D.-H. Noh, and J. Choi, "A study on the correlation between smart farm tomato production and growth data using multiple regression analysis," in *Proc. Symp. KICS*, pp. 763-764, Jun. 2022. (https://doi.org/10.7465/jkdi.2017.28.6.1427)

[5] J. Kim, S. Kwon, I. D. Ha, and M. H. Na, "Prediction of smart farm tomato harvest time: Comparison of machine learning and deep learning approaches," *J. Korean Data & Inf. Sci. Soc.*, vol. 33, no. 2, pp. 283-298, 2022. (https://doi.org/10.7465/jkdi.2022.33.2.283)

[6] S. Naoki, M. Hirokazu, N. Stephanie, Y. Satoshi, K. S. Takashi, N. Yo, and S. Kazuhito "3D measurement of an orchard using a LiDAR- camera system for use in digital twins," *The XX CIGR World Congress 2022*, Dec. 2022.

[7] "*Academic database search site 『DBpia』*," Retrieved Jan. 30, 2023, from https://www.dbpia.co.kr

[8] K. S. Park, Y. K. Yoo, H. J. Park, and J. E. Son, "Current status of korean smart farm research and technology development," *2022 Annu. Autumn Conf. Korea Soc. for Horticultural Sci. and Technol.*, pp. 51-52, Nov. 2022.

[9] S.-J. Kim and H. Yoe, "Trend and standardization of smart farm technology," *J. KICS*, vol. 47, no. 11, pp. 1965-1973, 2022. (https://doi.org/10.7840/kics.2022.47.11.1965)

[10] S. Kim, "Design and implementation of smartfarm integrated platform," *J. KICS*, vol. 46, no. 12, pp. 2403-2410, 2021. (https://doi.org/10.7840/kics.2021.46.12.2403)

[11] C.-H. Lee and S.-G. Hwang, "Design of environmental control system for smart farm cultivation," *J. KIIS*, vol. 32, no. 2, pp. 110-116, 2022. (https://doi.org/10.5391/JKIIS.2022.32.2.110)

[12] G. Park, K. Sim, J. Baek, S. Lee, and J.-I. Park, "Plant diseases detection algorithm in

smart farm phenomics system," *The Korean Inst. Broadcast and Media Eng. Summer Conf.*, Jun. 2022.

[13] K. Park, J. Hwang, G. Hong, and D. Suh, "A study on object detection with deep learning methods for smart farm image analysis," in *Proc. Symp. KICS*, pp. 1608-1609, Feb. 2022.

[14] H. Noh and Y. Lee, "Determinants of growth variables on smart farm tomato production," *The Soc. Convergence Knowledge Trans.*, vol. 8, no. 3, pp. 17-25, 2020. (https://doi.org/10.22716/sckt.2020.8.3.016)

[15] S. Hong, T. Park, J. Bang, and H. Kim, "A study on the prediction model for tomato production and growth using ConvLSTM," *The J. KIIT*, vol. 18, no. 1, pp. 1-10, 2020. (https://doi.org/10.14801/jkiit.2020.18.1.1)

[16] S.-Y. Lee, J. Kim, H.-R. Kim, J.-M. Yoo, and A-Y. Son, "A study on optimized prediction model selection of strawberry production for dpaas in distributed cloud computing," *J. KICS*, vol. 47, no. 3, pp. 539-550, 2022. (https://doi.org/10.7840/kics.2022.47.3.539)

[17] J.-H. Kim and J.-Y. Kim, "Comparative analysis of performance of BI-LSTM and GRU algorithm for predicting the number of Covid-19 confirmed cases," *J. KIICE*, vol. 26, no. 2, pp. 187-192, 2022. (http://doi.org/10.6109/jkiice.2022.26.2.187)

[18] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991*, Aug. 2015.

[19] "*Rural Development Administration - public use of smart farm excellent farms*," Retrieved Dec. 15, 2022, from https://www.data.go.kr /data/15042594/openapi.do

[20] R. Fifriani and P. W. Santosa, "Application of altman modified Z-Score to predict financial distress in the indonesian telecommunications industry," *J. Econ. and Business Aseanomics*, vol. 4, no. 1, pp. 23-35, 2019. (https://doi.org/10.33476/j.e.b.a.v4i1.1236)

[21] M. M. Ahsan, M. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, pp. 52, 2021. (https://doi.org/10.3390/technologies9030052)

[22] C. M. Andersen and R. Bro, "Variable selection in regression－a tutorial," *J. Chemometrics*, vol. 24, no. 11-12, pp. 728-737, 2010. (https://doi.org/10.1002/cem.1360)

[23] A. M. Olusegun, H. G. Dikko, and S. U. Gulumbe, "Identifying the limitation of stepwise selection for variable selection in regression analysis," *Am. J. Theoretical and Appl. Statistics*, vol. 4, no. 5, pp. 414-419, Sep. 2015. (https://doi.org/10.11648/j.ajtas.20150405.22)

[24] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surv.*, vol. 4, pp. 40-79, 2010. (https://doi.org/10.1214/09-SS054)

[25] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 281-305, 2012.

[26] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?," *Arguments against avoiding RMSE in the literature, Geosci. Model Dev.*, vol. 7, pp. 1247-1250. Jun. 2014. (https://doi.org/10.5194/gmdd-7-1525-2014)

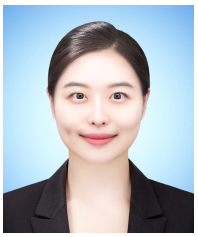
[27] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, no. e623, 2021. (https://doi.org/10.7717/peerj-cs.623)

이 세 연 (Se-Yun Lee)

2013년 2월 : 경희대학교 컴퓨터공학과 졸업
2015년 2월 : 경희대학교 컴퓨터공학과 석사
2020년 3월~현재 : 차세대융합기술연구원 연구원

<관심분야> 클라우드 컴퓨팅, 빅데이터, 가상화
[ORCID:0000-0003-1157-167X]

양 현 정 (Hyeonjeong Yang)



2023년 2월 : 덕성여자대학교 경영학·사회학과 졸업
<관심분야> 데이터분석, 데이터 시각화, 빅데이터

김 민 영 (Minyoung Kim)

2020년 3월~현재 : 숭실대학교 산업·정보시스템공학과 학사과정
<관심분야> 데이터마이닝, 데이터엔지니어링

김 준 경 (Junkyeong Kim)

2010년 2월 : 성균관대학교 고분자시스템공학과 졸업
2012년 2월 : 성균관대학교 u-City 공학과 석사
2018년 8월 : 성균관대학교 건설환경시스템공학과 박사
2019년 5월~현재 : 차세대융합기술연구원 선임연구원
<관심분야> IoT 센서, 건전성 모니터링(SHM)
[ORCID:0000-0002-6503-7950]

손 아 영 (A-Young Son)

2013년 2월 : 경희대학교 컴퓨터공학과 졸업
2015년 2월 : 경희대학교 컴퓨터공학과 석사
2020년 2월 : 경희대학교 컴퓨터공학과 박사
2020년 4월~2020년 12월 : 한국건설기술연구원 연구원
2021년 1월~2022년 11월 : 차세대융합기술연구원 선임연구원
2022년 12월~현재 : 한국과학기술정보연구원 선임연구원
<관심분야> 클라우드 컴퓨팅, AI, 빅데이터
[ORCID:0000-0002-8291-6033]

홍 성 훈 (Seonghun Hong)

2007년 2월 : 한양대학교 전자전기컴퓨터공학부 졸업
2017년 2월 : 한양대학교 전기공학과 박사
2007년 3월~2017년 2월 : 한국과학기술연구원 로봇·미디어 연구소 학연 박사과정
2017년 3월~2020년 12월 : 두산인프라코어 기술원 선임연구원
2021년 1월~현재 : 차세대융합기술연구원 선임연구원
2022년 12월~현재 : ISO/TC 299 - Robotics/WG 1,4,6 Committee Member
<관심분야> 시스템공학, 메카트로닉스, 로봇공학
[ORCID:0000-0002-7888-6257]