Journal of KIIT. Vol. 23, No. 1, pp. 21-29, Jan. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.1.21

토마토 수확량 및 LAI 예측을 위한 순환 신경망 기반 딥러닝 모델들의 성능 비교 연구

노혜민*, 장기성**, 이지현***

A Comparative Study on the Performance of RNN-based Deep Learning Models for Tomato Yield and LAI Prediction

Hyemin Noh*, Giseong Jang**, and Jihyun Lee***

요 약

정밀 농업 시스템에서 농작물의 수확량 예측 서비스가 중요해지면서 인공지능 모델의 활용이 요구되고 있으나 어떤 인공지능 모델이 농작물 수확량 예측에 적합한지는 아직 확인지 않았다. 본 연구에서는 순환 신경망 기반의 딥러닝 모델인 RNN, LSTM, Bi-LSTM, GRU, Bi-GRU를 대상으로 토마토 작물의 수확량과 LAI(엽면적지수)의 예측 성능을 비교한다. 비교 실험을 위해 3년간 각 2주 간격으로 수집된 토마토 재배 환경 및 생육 데이터를 활용하였으며, 결측치 처리, 데이터 변환, 요약 변수 생성, 업샘플링, 스케일링 등의 전처리 과정을 거쳐 총 943개의 데이터를 확보하였다. 실험 결과, 표준 스케일러-LSTM과 거듭제곱 변환 스케일러-Bi-LSTM 조합이 수확량과 LAI의 예측에서 가장 작은 오차를 보여 토마토 작물의 수확량 예측에 적합한 것으로 확인되었다.

Abstract

As yield prediction services become increasingly important in precision agriculture systems, the use of AI models is in demand. However, it remains unclear which AI model is best suited for crop yield prediction. This study compares the prediction performance of recurrent neural network-based deep learnin models－RNN, LSTM, Bi-LSTM, GRU, and Bi-GRU for forecasting tomato crop yield and Leaf Area Index(LAI). For the comparative experiment, tomato cultivation environment and growth data collected at two-week intervals over three years were used, and a total of 943 data points were obtained after preprocessing steps, including handling missing values, data transformation, feature engineering, upsampling, and scaling. The experimental results showed that the Standard Scaler-LSTM and Power Transformer-Bi-LSTM combinations had the smallest errors in predicting yield and LAI, indicating their suitability for forecasting tomato crop yield.

Keywords

precision agriculture, crop yield prediction, LAI prediction, recursive neural network, tomato plant

* 전북대학교 소프트웨어공학과 강의초빙교수
- ORCID: https://orcid.org/0000-0003-1150-3570
** ㈜에스에스엘 과장
- ORCID: https://orcid.org/0009-0001-6253-0774
*** 전북대학교 소프트웨어공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0003-4512-806X

· Received: Oct. 21, 2024, Revised: Nov. 18, 2024, Accepted: Nov. 21, 2024
· Corresponding Author: Jihyun lee
Dept. of Software Engineering, Jeonbuk National University
Tel.: +82-63-270-4860, Email: jihyun30@jbnu.ac.kr

Ⅰ. 서 론

최근 기후 변화, 재배 환경의 변화, 에너지 가격 상승 등으로 인해 농업에서 적절한 농작물 생산량을 유지하는 것이 점점 더 중요해지고 있다. 이러한 적정 농작물 생산량 유지를 위해 농작물 수확량 예측은 필수적이다. 최근에는 정밀농업에서 인공지능 모델을 적용해 농작물의 수확량을 예측하는 연구들이 진행되어 왔다[1]. 작물 수확량 예측을 위해 머신러닝과 딥러닝 모델의 성능을 비교한 여러 연구에서 LSTM(Long-Short Term Memory)가 가장 성능이 좋은 것으로 보고되었다[2]-[7]. 그러나 기존 연구들은 시계열 데이터를 기반으로 한 예측에서 성능이 좋은 것으로 알려진 모델들을 대상으로 성능을 비교하지는 않았다.

이에 본 연구에서는 스마트팜 농장에서 수집한 토마토의 재배 환경 및 생육 데이터를 활용하여 시계열 데이터를 기반으로 한 예측에서 성능이 좋은 것으로 알려진 순환 신경망 기반의 딥러닝 모델들의 수확량과 LAI(Leaf Area Index) 예측 성능을 비교한다. 순환 신경망 기반의 딥러닝 모델은 RNN(Recurrent Neural Network), GRU(Gated Recurrent Unit), Bi-GRU(Bidirectional GRU), LSTM, Bi-LSTM(Bidirectional LSTM), 를 선정하여 비교 실험을 수행한다.

Ⅱ. 관련 연구 및 배경지식

2.1 인공지능 모델을 활용한 수확량 예측

인공지능 기술은 어떤 작물을 재배할지를 결정하고, 작물의 생장 과정에서 생장 환경을 어떻게 조정할지와 같은 의사결정을 포함하여, 작물 수확량 예측에 활발히 적용되고 있다. 특히, CNN(Convolutional Neural Networks), LSTM, DNN (Deep Neural Networks)가 가장 널리 사용되고 있다[1]. ADENIYI는 토양과 외부 기후 요인을 입력으로 딥러닝 모델의 성능을 비교 분석한 결과 LSTM이 가장 높은 정확도를 보였다고 보고하였다[2].

관련 연구 [3]-[8]은 환경 및 생육 데이터를 토대로 토마토 수확량을 예측한 연구이다. Alwis는 LSTM과 어텐션 점수 메커니즘(Attention score mechanism)을 적용한 Duo Attention-LSTM 모델을 제안하고, 이 모델을 이용하여 토마토 수확량 예측과 수확량에 영향을 미치는 요인 분석을 진행하였다[3]. 이 연구는 Duo Attention-LSTM이 LSTM 보다 예측 성능이 우수함을 확인했으며, 환경 요인으로는 CO2 농도가 생육 요인으로는 개화 속도가 수확량에 주요하게 영향을 미치는 요인으로 제시하고 있다. 그러나 이 연구는 생육 요인으로 개화 속도, 착과 속도, 엽수, LAI를 사용하고 있는데 엽수가 LAI에 포함된 요인임을 고려하면 사용한 생육 요인은 2개뿐이라는 점에서 한계가 있다.

S. Kang et al.[4]은 Dual Attention LSTM과 LSTM, Attention LSTM의 토마토 수확량 예측 성능을 비교하고 있다. 이 연구는 Dual Attention LSTM이 가장 좋은 예측 성능을 보인 것으로 보고하고 있지만, 환경 요인만을 예측에 활용하였다는 점에서 본 연구와 차이가 있다. 게다가 이 연구는 모든 환경 요인이 생산량 예측에 비슷한 정도로 중요하다고 보고하고 있어 [3]의 결과와 배치될 뿐만 아니라 MSE(Mean Squared Error)가 0.678로 예측 오차 또한 상당히 커서 실험 결과를 신뢰하기 어렵다.

S. Hong et al.[5]의 연구는 다중 회귀 분석, 랜덤 포레스트(Random forest), 콘볼루션 레이어를 추가한 LSTM인 딥러닝 ConvLSTM(Convolution LSTM)의 생산량 및 생장량 예측 성능을 비교한 결과 딥러닝 ConvLSTM이 가장 우수한 성능을 보였다고 보고하였다. 이 연구는 본 연구와 유사한 환경 및 생육 요인을 활용하고 있으나 생산량 예측의 MAE(Mean Absolute Error)와 RMSE가 각 3.588과 4.559로 높다.

A. Alhnaity et al.[6]의 온실 환경의 토마토 수확량 예측 연구에서도 LSTM이 랜덤 포레스트나 서포트 벡터 회귀(Support vector regression) 보다 더 우수한 예측 성능이 보였다. 이 연구의 MSE, RMSE, MAE는 각각 0.002, 0.047, 0.03으로 앞서 소개된 연구들 보다 낮다. 그러나 예측에 사용한 환경 요인이 CO2, 습도, 온실 내부 온도, 외부 온도, 광량 등 외부 환경 요인까지를 포함하고 있고, 생육 요인으로는 줄기 굵기만을 사용했다는 점에서 본 연구와 차이가 있다.

S.-W. Kim et al.[7]은 릿지 회귀(Ridge regression), 랜덤 포레스트, XGBoost(Extreme Gradient Boosting)를 사용하여 토마토 수확량을 예측하였고 XGBoost가 가장 좋은 성능을 보였다. 그러나 이 연구는 생육 요인만을 예측에 활용하였다는 점에서 본 연구와 차이가 있으며 MAE와 RMSE가 각각 0.233과 0.817로 높다. 마지막으로 Belouz는 인공지능 모델을 활용해 토마토 수확량을 예측했지만, 연구의 목적이 에너지 사용 패턴을 추적하는 것이다[8].

본 연구는 재배 환경 및 생육 요인을 입력으로 토마토 수확량 예측을 한다는 점에서 이전 연구들과 유사하지만, 온실 내부의 환경 요인만을 입력으로 사용하면서 줄기 굵기 외에도 주차, 생장 길이, 화방 높이, 개화군, 착과군 등 더 많은 생육 요인을 함께 고려하여 예측한다는 점에서 차별화된다. 또한 대부분의 기존 연구들에서 좋은 성능을 보인 LSTM이 속한 순환 신경망 기반 딥러닝 모델들만을 선택하여 성능을 비교한다는 점에서 다르다.

2.2 순환 신경망 기반 딥러닝 모델

농작물 재배 환경 및 생육 데이터는 시계열 특성을 가지기 때문에 예측에는 순환 신경망 기반의 딥러닝 모델이 적합하다. 이 절에서는 기본 순환 신경망을 포함하여 순환 신경망 기반 딥러닝 모델에는 다음과 같은 모델들이 있다.

2.2.1 RNN(Recurrent Neural Network)

첫 번째 성능 비교 모델인 RNN은 자연어 처리, 음성 인식, 시계열 데이터 등에서 사용된다[9]. RNN은 순서가 있는 시퀀스 데이터를 다루며, 이전 시점의 정보를 활용하여 현재 출력을 생성하는 특성이 있다. 이러한 특성으로 인해 순차적인 데이터를 효과적으로 처리할 수 있지만, 시퀀스가 길어질수록 장기 의존성 문제가 발생하여 이전의 정보의 영향력이 줄어드는 단점이 있다.

2.2.2 LSTM(Long-Short Term Memory)

LSTM은 RNN의 장기 의존성 문제를 해결하기 위해 개발된 모델로, 셀 상태(Cell state)라는 개념을 도입하여 정보를 장기간 기억한다. 이를 위해 입력 게이트(Input gate), 망각 게이트(Forget gate), 출력 게이트(Output gate)라는 세 가지 게이트를 사용하여 셀 상태를 어떻게 업데이트한다. 이러한 구조 덕분에 LSTM은 장기적인 의존성을 가진 데이터를 효과적으로 학습할 수 있다[10]. 그러나 LSTM은 단방향(순방향)으로만 정보를 처리하므로 과거 정보만을 활용한다.

2.2.3 Bi-LSTM(Bidirectional LSTM)

Bi-LSTM은 순방향과 역방향 두 방향에서 정보를 처리하며, 두 방향의 예측 결과를 이어 붙이기(Concatenation)하거나 결합하여 최종 출력을 생성한다. LTSM에 역방향 흐름을 추가하여 미래 정보도 활용할 수 있도록 한 모델이다[11][12]. LSTM과 Bi-LSTM은 일반적인 RNN 보다 더 나은 성능을 보이지만, 계산량이 많다는 단점이 있다.

2.2.4 GRU(Gated Recurrent Unit)

이 모델은 Bi-LSTM 보다 구조가 단순하여 계산 속도가 빠르면서 유사한 성능을 제공하는 모델이다. GRU는 LSTM의 입력 게이트, 삭제 게이트, 출력 게이트 대신 리셋 게이트(Reset gate)와 업데이트 게이트(Update gate)를 사용한다[13][14]. 리셋 게이트는 과거 정보의 양을 조절하는 역할을 한다.

2.2.5 Bi-GRU(Bidirectional GRU)

GRU의 단방향 처리 한계를 극복하기 위해 개발된 모델로, Bi-LSTM과 유사하게 양방향에서 정보를 처리한다.

III. 데이터 셋 확보 및 전처리

3.1 환경·생육 데이터 셋

비교 연구는 1주일간의 성장 변화가 뚜렷하고 과육의 개수가 많은 특징을 갖는 토마토를 비교 연구에 사용할 작물로 선택하였다. 데이터 셋은 농림수산식품교육문화정보원이 운영하는 스마트팜 데이터

마트에 등록된 토마토의 재배 환경과 생육 데이터 셋 중, 비교적 최근인 2020년부터 2022년까지의 데이터를 품종 구분 없이 추출하였다[15]. 동일한 재배 조건인 '비닐 온실'과 '양액' 사용을 기준으로 데이터를 추출한 결과, 표 1과 같이 5개 지역에 위치한 8개 농가의 환경 및 생육 데이터를 최종 확보하였다.

표 1. 년도 및 지역 별 데이터 셋의 규모
Table 1. Size of integrated dataset by year and region

Region	Year	Size of dataset
Yeoju, Gyeonggi Province	2021	134
Sacheon, Gyeongsangnam-do	2020	184
	2021	99
	2022	120
Jinju, Gyeongsangnam-do	2021	198
Yeongam, Jeollanam-do	2021	29
	2022	187
Jeongeup, Jeollabuk-do	2021	208
Total		1,159

3.2 데이터 전처리

데이터 전처리는 (1) 특징 선택 및 결측치 처리, (2) 데이터 변환 및 요약 변수 생성, (3) 시퀀스 길이 통일 및 데이터 균형화, (4) 데이터 스케일링 순서로 진행하였다.

3.2.1 특징 선택 및 결측치 처리

환경 데이터는 결측치가 많아 보정이 어려운 특징(Feature)들이 다수 있었다. 이러한 특징들을 제외하고 내부 온도(℃), 내부 습도(%), 내부 CO2(ppm)의 3가지 특징을 선택하였다. 생육 데이터는 주차, 생장 길이(mm), 화방 높이(mm), 줄기 굵기(mm), 개화군(점), 착과군(점)을 선택하였다. 생장 길이 데이터에서 발견된 결측치는 평균 생장 길이로 보정하였다. 수확량 정보로는 데이터 셋에서 열매 수(개)를 추출하여 활용하였고, LAI는 데이터 셋에 포함된 보고서가 제공하는 수식을 기반으로 산출하였다. 이 수식은 식 (1)과 같이 생육 데이터의 엽수(개), 엽장(cm), 엽폭(cm), 정식 주수와 재배 면적(m2)을 활용한다.

$$LAI = 엽수 \times 엽장 \times 엽폭 \times 재식\ 밀도 \times 60\% \quad (1)$$

단, 재식 밀도는 정식 주수/㎡임

표 2는 선택한 환경 및 생육 관련 특징과 단위를 보여준다.

표 2. 실험에 사용된 토마토 작물의 환경 및 생육 특징
Table 2. Environmental and growth features of tomato crops

Category	Feature	Unit
Environmental information	Air temperature inside the greenhouse	℃
	Air relative humidity inside the greenhouse	%
	Air CO_2 concentration inside the greenhouse	ppm
Crop growth information	Weeks after planting	week
	Growth length	mm
	Inflorescence height	mm
	Stem thickness	mm
	Flowering body	point
	Fruiting cluster	point
	Number of fruits	-
	LAI	-
Basic data	Year, location, data collection date	

3.2.2 데이터 변환 및 요약 변수 생성

환경 데이터와 생육 데이터는 일 단위로 주기를 변환하는 작업을 진행하였다.

환경 데이터의 경우, 긴 시계열 데이터 셋을 압축하여 학습 데이터로 사용하기 위해 시계열 데이터에서 패턴을 추출하는 시계열 데이터 특징 추출(Time series feature extraction) 방법을 활용하였다. 시간 단위 데이터가 일 단위로 변환되는 과정에서 시계열 특성을 유지하기 위해 최소(_min), 1/4분위수(_25), 평균(_mean), 중앙(_median), 3/4분위수(_75), 최대값(_max)으로 요약 변수를 생성하였다. 내부 온도, 내부 습도, 내부 CO2 각각에 대해 요약 변수가 생성되어 환경 데이터는 총 18개의 특징으로 요약되었다. 생육 데이터의 경우, 주 단위 데이터를 일 단위로 변환하기 위해 농장 별로 선형 보간법을 적용하였다. 1주일 단위로 수집되어야 하는 데이터가 2주일 이상의 간격으로 수집된 경우에는 별도의 처리 없이 선형 보간법을 통해 보정하였다. 그 결과 데이터는 총 26개의 특징으로 조정되었다.

재배 환경 및 생육 데이터의 특징들은 일반적으로 서로 선형적 관계를 가진다. 그렇지만 특징들 간 비선형적 패턴이 없다고 확신할 수는 없다. 따라서 조정된 1,159개의 데이터를 입력으로 26개 특징들 간 선형적 상관관계와 비선형적 상관관계를 분석하였다. 그림 1은 특징 벡터들 간의 피어슨 상관분석 결과이다.

상관관계 분석 결과의 통계적 유의성 검증을 위해 다음과 같은 귀무가설과 대립가설을 기반으로 p-value 값을 도출하였다.

- 귀무가설(H_0): 두 변수 사이에 상관관계가 없다.
- 대립가설(H_1): 두 변수 사이에 상관관계가 있다.

그림 2는 도출한 p-value 값을 히트맵으로 표현한 결과이다. 대부분의 p-value 값이 0.05보다 작으므로 귀무가설을 기각하고 대립가설을 채택할 수 있다.

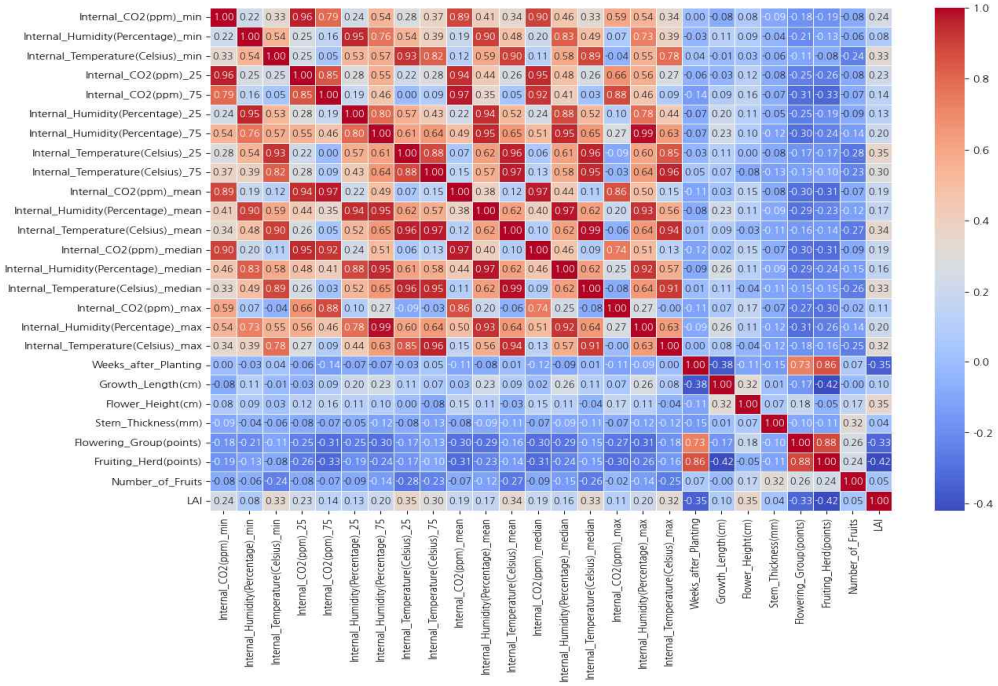

그림 1. 특징 벡터들 간의 상관관계
Fig. 1. Correlation of feature vectors

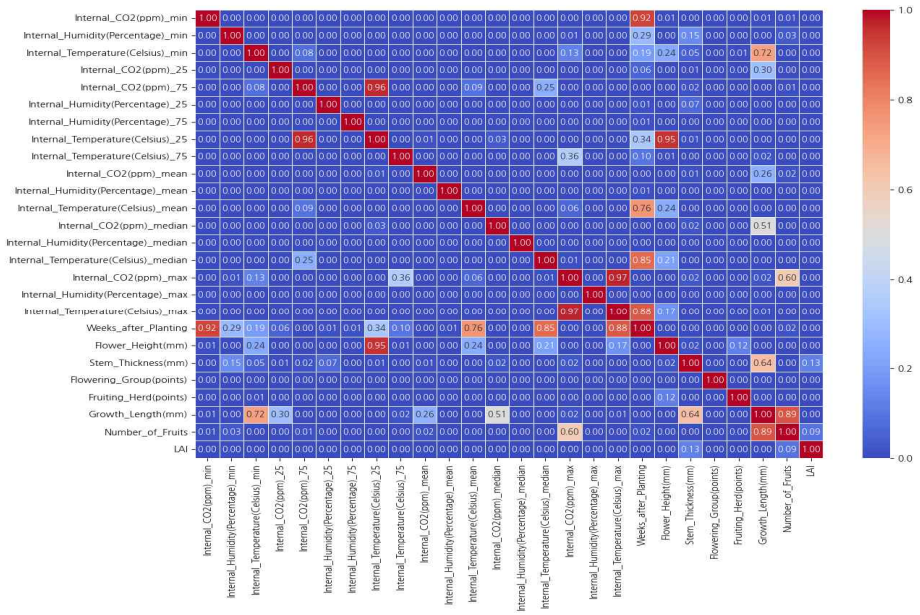

그림 2. 상관관계 유의성 분석 결과 (p-value 히트맵)
Fig. 2. Correlation significance analysis results (p-value heatmap)

재배 환경 정보의 특징들(온도, 습도, CO2) 간 관계는 대부분 양의 상관관계를 보인 반면, 생육 정보의 특징들 간 관계는 대부분 음의 상관관계가 관찰되었다. 수확량과 LAI는 다른 변수들과 약한 상관관계를 보였으며, 동일 특징에서 파생된 요약 변수들 간에만 강한 상관관계가 관찰되었다. 이 결과는 스피어만 상관분석에서도 동일하였다. 따라서 특징들 간에는 선형적 상관관계만 존재한다고 볼 수 있어 모든 특징을 실험에 활용하였다.

3.2.3 시퀀스 길이 통일 및 데이터 균형화

시퀀스 길이는 입력과 출력 모두 14일(2주)로 설정하였으며 이 과정에서 데이터 수량이 줄었다. 이를 해결하기 위해 생육 데이터를 일 단위로 변환하여 양을 늘리는 업샘플링 작업을 진행하였으며 총 943개의 데이터를 확보하였다.

3.2.4 데이터 스케일링

재배 환경 데이터와 생육 데이터는 값의 범위가 서로 다르므로 데이터 표준화가 필요하다. 그러나 주어진 데이터 셋의 특성과 모델에 가장 적합한 스케일링 방법을 사전에 알 수 없다. 따라서 본 연구는 최소-최대 스케일러(Min max scaler), 표준 스케일러(Standard scaler), 강건 스케일러(Robust scaler), 거듭제곱 변환 스케일러(Power transformer)[16][17], 분위수 변환 스케일러(Quantile transformer)[18]를 각 모델에 모두 적용하였다.

또한, 데이터 전처리 과정에서 사용한 결측치 처리 방법과 선형 보간법 적용의 타당성을 검증하기 위해, 보간 전후 데이터 차이를 시각적으로 표현하여 데이터의 본래 특성이 유지되는지 확인하였다. 그 결과, 보간 전후의 그래프는 주요 피크와 데이터 패턴이 일치하였으며, 보간된 구간에서도 이탈이나 급격한 변동이 없었다.

Ⅳ. 성능 평가 시험 및 결과

4.1 실험 환경 및 파라미터

순환 신경망 모델들의 성능을 평가에 사용한 하드웨어와 소프트웨어 환경은 표 3과 같다.

표 3. 실험 환경
Table 3. Experiment environment

Environment		Specifications
Physical environment	OS	Ubuntu 22.04
	CPU	AMD Ryzen 7 5800X
	GPU	Nvidia GeForce RTX 3070
	RAM	16 GB
IDE (Integrated Development Environment)	Python	3.8.10
	IPython	8.11.0
	numpy	1.23.5
	scipy	1.10.1
	keras	2.12.0
	sklearn	1.3.0
	pandas	2.0.3
	tensorflow	2.12.0

손실 함수로는 MAE를, 최적화 함수로는 Adam(Adaptive moment estimation)을, 학습률 스케줄러는 EarlyStopping과 ReduceLROnPlateau를 각각 사용하였다. 하이퍼 파라미터는 별도의 튜닝 작업 없이 다양한 테스트를 수행한 후 가장 높은 성능을 보인 값을 선택하였다. 표 4는 유형 별 파라미터들과 설정 값을 보여준다.

표 4. 모델에 적용한 하이퍼 파라미터
Table 4. Hyper parameters applied to models

Types	Parameters and values
General hyper-parameters	Batch size: 8 Learning rate: 0.0001 Epochs: 1000
EarlyStopping hyper-parameters	monitor: val_loss min_delta: 0.0001 patience: 50 mode: auto
ReduceLROnPlateau hyper-parameters	monitor: val_loss factor: 0.3 patience: 25 min_lr: 0.000005

학습, 검증, 시험 셋은 총 943개의 데이터를 8:1:1 비율로 분할하여 사용하였으며, 성능 비교를 위한 평가 지표로는 MAE, RMSE, SMAPE(Symmetric Mean Absolute Percentage Error)를 사용하였다.

4.2 결과 및 분석

각 스케일러-모델 조합을 테스트한 결과는 표 5와 같다.

표 5. 스케일러 별 모델의 예측 오차
Table 5. Prediction errors of models based on different scalers

Models \ Scalers		MinMax			Standard			Robust			Power transformer			Quantile transformer		
		MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE	MAE	RMSE	SMAPE
RNN	# fruits	0.3	0.41	5.22	0.14	0.2	3.53	***0.13***	***0.18***	***3.39***	0.14	0.2	3.7	0.25	0.58	3.95
	LAI	0.06	0.08	5.41	0.03	0.05	2.87	0.03	0.04	2.51	***0.03***	***0.04***	***2.08***	0.04	0.08	3.2
GRU	# fruits	0.17	0.26	3.71	0.05	0.08	2.2	***0.04***	***0.08***	***2.1***	0.07	0.1	2.87	0.13	0.33	2.68
	LAI	0.04	0.05	3.33	0.08	0.11	2.89	0.07	0.11	2.8	***0.02***	***0.03***	***1.24***	0.03	0.05	1.96
Bi-GRU	# fruits	0.2	0.3	3.87	0.02	0.03	1.51	***0.02***	***0.03***	***1.4***	0.08	0.11	2.92	0.12	0.24	2.11
	LAI	0.04	0.06	3.72	0.08	0.12	2.89	0.04	0.08	2.12	***0.02***	***0.03***	***1.29***	0.03	0.05	1.92
LSTM	# fruits	0.19	0.28	3.81	***0.02***	***0.03***	***1.58***	0.07	0.11	2.86	0.07	0.1	2.79	0.08	0.15	1.42
	LAI	0.04	0.06	3.85	***0.06***	***0.09***	***2.77***	***0.02***	***0.02***	***1.38***	***0.02***	***0.03***	***1.21***	***0.02***	***0.04***	***1.53***
Bi-LSTM	# fruits	0.13	0.2	3.32	***0.01***	***0.02***	***1.27***	0.04	0.07	2.08	***0.06***	***0.1***	***2.75***	0.09	0.17	1.47
	LAI	0.03	0.04	2.44	0.07	0.11	2.84	0.06	0.09	2.81	***0.01***	***0.03***	***1.18***	0.02	0.04	1.54

먼저 모델 별로 수확량 예측에 적합한 스케일러를 살펴보면, RNN, GRU, Bi-GRU는 강건 스케일러와의 조합에서, LSTM과 Bi-LSTM은 표준 스케일러와의 조합에서 좋은 예측 성능을 보였다. LAI에서는 모든 모델이 거듭제곱 변환 스케일러와의 조합에서 좋은 성능을 보였다. LSTM은 강건 스케일러와 분위수 변환 스케일러에서도 좋은 예측 성능을 보였다.

두 번째로, 좋은 예측 성능을 보인 스케일러-모델 조합에서 수확량과 LAI 예측 결과를 살펴보면, 수확량은 표준 스케일러-Bi-LSTM 조합이, LAI는 거듭제곱 변환 스케일러-Bi-LSTM 조합이 각각 가장 좋은 성능을 보였다.

세 번째로, 수확량과 LAI를 함께 고려해서 예측 결과를 표준 스케일러-LSTM과 거듭제곱 변환 스케일러-Bi-LSTM 조합의 예측 성능이 우수했다.

네 번째로, 평가지표를 보면, 모든 실험 결과에서 MAE와 RMSE는 값이 작고 SMAPE 값은 이 두 값에 비하면 상대적으로 크다. 이는 예측 값과 실제 값 간의 차이가 매우 작았기 때문인 것으로 분석되었다.

마지막으로, LSTM과 Bi-LSTM의 예측 결과를 스케일러에 관계없이 살펴보면, LAI의 MAE와 RMSE 값의 오차는 0.09~0.01로 비교적 낮은 반면, 수확량은 0.28~0.01로 LAI 보다 오차 범위가 다소 넓다. 이 결과의 원인은 두 가지로 설명할 수 있다. 첫째, 수확량 데이터가 LAI보다 이상치가 더 많다는 점이다. 둘째, 수확량은 실험에서 사용한 특징 이외에도 다른 요인에 영향을 받을 수 있다는 점이다.

Ⅴ. 결론 및 향후 연구

본 연구에서는 토마토 작물의 수확량 및 LAI 예측에 적합한 스케일러와 딥러닝 모델을 찾기 위해 5개의 스케일러와 5개의 순환 신경망 기반 딥러닝 모델을 선정하여 총 25개의 스케일러-모델의 조합으로 비교 실험을 진행하였다. 실험은 토마토 재배 환경 및 생육 데이터 943개를 이용하여 수행되었다. 그 결과, 스케일러에 따라 일부 차이가 있었으나 표준 스케일러를 적용한 LSTM과 거듭제곱 변환 스케일러를 적용한 Bi-LSTM이 다른 조합에 비해 예측 오차가 낮았다. 그러나 수확량과 LAI 모두에서 가장 우수한 예측 성능을 보인 단일 스케일러-모델 조합은 없었다. 본 실험을 통해, 작물 재배 환경 및 생육 데이터의 스케일링에는 표준 스케일러와 거듭제곱 변환 스케일러가 적합하며, 수확량 예측에는 LSTM과 Bi-LSTM이 효과적임을 확인하였다.

향후 연구에서는 먼저, 본 실험이 토마토의 품종 정보가 반영되지 않은 데이터 셋을 기반으로 진행되었으므로, 품종 정보가 결과에 미치는 영향을 확인하기 위해 동일 품종 데이터를 활용한 추가 실험이 필요하다. 또한, 환경 및 생육 특징을 추가하거나 변경한 뒤 비교 실험을 수행하여 수확량 예측에서 오차 범위가 크게 나타난 원인을 규명하고, 수확량 예측에 영향을 미치는 다른 요인을 탐색할 계획이다.

마지막으로, 현재 운영 중인 정밀농업 지원 시스템에 예측 성능이 우수한 스케일러-모델 조합을 적용하여, 현재 시스템이 실시간으로 수집하는 환경 데이터와 주기적으로 관리자가 수집하여 기록하는 생육 데이터를 기반으로 수확량과 LAI를 예측하는 서비스를 새롭게 추가할 예정이다.

References

[1] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review", Computers and Electronics in Agriculture, Vol. 177, pp. 1-18, Oct. 2020. https://doi.org/10.1016/j.compag.2020.105709.

[2] J. K. Adeniyi, T. T. Adeniyi, S. A. Ajagbe, E. A. Adeniyi, O. Aiyeniko, and M. O. Adigun, "A comparative analysis of the performance of deep learning techniques in precision farming using soil and climate factors", Procedia Computer Science, Vol. 235, pp. 2812-2821, May 2024. https://doi.org/10.1016/j.procs.2024.04.266.

[3] S. D. Alwis, Y. Zhang, M. Na, and G. Li, "Duo attention with deep learning on tomato yield prediction and factor interpretation", PRICAI 2019: Trends in Artificial Intelligence, Vol. 11672, pp. 704-715, Aug. 2019. https://doi.org/10.1007/978-3-030-29894-4_56.

[4] S. Kang, K. Cho, and M. H. Na, "Forecasting crop yield using encoder-decoder model with attention", Journal of Korean Society Quality Management, Vol. 49, No. 4, pp. 569-579, Dec. 2021. http://doi.org/10.7469/JKSQM.2021.49.4.569.

[5] S. Hong, T. Park, J. Bang, and H. Kim, "A study on the prediction model for tomato production and growth using ConvLSTM", Journal of KIIT, Vol. 18, No. 1, pp. 1-10, Jan. 2020. http://doi.org/10.14801/jkiit.2020.18.1.1.

[6] B. Alhnaity, S. Pearson, G. Leontidis, and S. Kollias, "Using deep learning to predict plant growth and yield in greenhouse environments", Acta Hortic, Vol. 1296, pp. 425-432, Nov. 2020. https://doi.org/10.17660/ActaHortic.2020.1296.55.

[7] S.-W. Kim and Y. Kim, "A study on the application of machine learning algorithm to predict crop production", Journal of the Korea Academia-Industrial cooperation Society, Vol. 22, No. 7 pp. 403-408, Jul. 2021.

[8] K. Belouz, A. Nourani, S. Zereg, and A. Bencheikh, "Prediction of greenhouse tomato yield using artificial neural networks combined with sensitivity analysis", Scientia Horticulturae, Vol. 293, pp. 1-8, Feb. 2022. https://doi.org/10.1016/j.scienta.2021.110666.

[9] R. J. Williams, G. E. Hinton, and D. E. Rumelhart, "Learning representations by back-propagating errors", Nature, Vol. 323, No. 6088, pp. 533-536, Oct. 1986. https://doi.org/10.1038/323533a0.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Computation, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997.

[11] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", Neural Networks, Vol. 18, No. 5-6, pp. 602-610, Jul.-Aug. 2005. https://doi.org/10.1016/j.neunet.2005.06.042.

[12] T. Thireou and M. Reczko, "Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No. 3, pp. 441-446, Jul.-Sep. 2007. https://doi.org/10.1109/tcbb.2007.1015.

[13] J. Heck and F. M. Salem, "Simplified minimal gated unit variations for recurrent neural networks", 2017 IEEE 60th International Midwest Symposium on Circuits and Systems(MWSCAS), Boston, MA, USA, pp. 1593-1596, Aug. 2017. https://doi.org/10.1109/MWSCAS.2017.8053242.

[14] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks", 2017 IEEE

60th International Midwest Symposium on Circuits and Systems(MWSCAS), Boston, MA, USA, pp. 1597-1600, Aug. 2017. https://doi.org/10.1109/MWSCAS.2017.8053243.

[15] Smarfarm Datamart, https://data.smartfarmkorea.net. [accessed: Nov. 08, 2024]

[16] G. E. P. Box and D. R. Cox, "An analysis of transformations", Journal of the Royal Statistical Society, Vol. 26, No. 2, pp. 211-243, Jul. 1964. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

[17] J. Johnston, "Econometric Methods", 3rd Ed., New York: McGraw-Hill, pp. 61-74, Jan. 1984.

[18] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry", Biometrika, Vol. 87, No. 4, pp. 954-959, Dec. 2000. https://doi.org/10.1093/biomet/87.4.954.

저자소개

노 혜 민 (Hyemin Noh)

2000년 2월 : 전북대학교 컴퓨터과학과(학사)
2002년 2월 : 전북대학교 전산통계학과(석사)
2006년 2월 : 전북대학교 컴퓨터통계정보학과(박사)
2011년 3월 ~ 현재 : 전북대학교 소프트웨어공학과 강의초빙교수
관심분야 : 정밀 농업, 빅데이터 분석

장 기 성 (Giseong Jang)

2020년 8월 : 전북대학교 소프트웨어공학과(학사)
2020년 1월 ~ 현재 : ㈜에스에스엘 과장
관심분야 : 데이터 분석, 머신러닝, 딥러닝

이 지 현 (Jihyun Lee)

1993년 8월 : 전북대학교 정보통신공학과(학사)
2000년 2월 : 전북대학교 전자계산교육(석사)
2005년 2월 : 전북대학교 컴퓨터과학과(박사)
2016년 3월 ~ 현재 : 전북대학교 소프트웨어공학과 교수
관심분야 : 소프트웨어제품라인, 아키텍처 재구축