스마트팜 토마토의 생산량 예측 모델을 위한 최적 환경 변수 선정 연구

(A Study on Optimal Environmental Variable Selection for Tomato Yield Prediction Models in Smartfarm)

김진만*, 홍아름**
(Jinman Kim, Ahreum Hong)

요 약

기후 변화와 노동력 부족 등 농업의 구조적 위기에 대응하기 위한 대안으로, 4차 산업혁명 기술을 농업에 접목한 스마트팜(smartfarm)이 빠르게 확산되고 있다. 그러나 현실적으로 IoT 기반 장비의 도입 비용과 유지·관리의 복잡성으로 인해, 실제 농업인들이 활용하기에는 한계가 존재한다. 본 연구는 국내 스마트팜 온실에서 대중적으로 보급된 시설 및 장비의 현황을 토대로, 토마토 생산량 예측을 위한 10개의 주요 환경 데이터 항목을 선정했다, 그리고 랜덤포레스트(RandomForest), 장단기기억모델(LSTM), 게이트순환유닛(GRU) 세 가지 모델을 적용하여 각 변수의 영향을 비교·분석하였다. 모델의 성능 평가는 RMSE와 결정계수(R^2)를 활용하였으며, 분석 결과 랜덤포레스트 모델이 4개의 변수로 학습하였을 때 가장 높은 성능을 보였다. 전반적으로 3~6개의 변수를 입력으로 사용한 경우 모델별 예측력이 향상되는 경향을 보였으며, 생산량 예측과의 연관성이 높은 주요 변수로는 내부 CO_2 농도, 외부 온도, 내부 온도, 공급 EC, 누적일사량 등이 도출되었다. 본 연구의 결과는 향후 농업 현장에서 스마트팜 고도화를 위한 시설 및 장비 도입 의사결정시, 과학적 근거자료로 활용될 수 있을 것으로 기대된다.

■ 중심어 : 스마트팜 ; 토마토 ; 정밀농업 ; 랜덤포레스트

Abstract

As an alternative to address the structural crisis in agriculture caused by climate change and labor shortages, smartfarm that integrate Fourth Industrial Revolution technologies are rapidly expanding. However, in reality, the high costs of IoT-based equipment installation and the complexity of maintenance and management pose significant barriers to practical adoption by farmers. This study selected 10 key environmental data variables for tomato yield prediction based on commonly deployed facilities and equipment in domestic smartfarm greenhouses. Three models—RandomForest, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU)—were applied to compare and analyze the impact of each variable. Model performance was evaluated using Root Mean Square Error (RMSE) and the coefficient of determination (R^2). The analysis revealed that the RandomForest model achieved the highest performance when trained with four variables. Overall, models showed improved predictive power when using 3 to 6 variables as inputs. Key variables highly correlated with yield prediction included internal CO_2 concentration, external temperature, internal temperature, supply EC, and cumulative solar radiation. The findings of this study are expected to serve as scientific evidence for decision-making regarding facility and equipment adoption in advancing smart farm implementation in agricultural settings.

■ keywords : Smartfarm ; Tomato ; Precision Agriculture ; RandomForest

Ⅰ. 서 론

기후 위기 및 전쟁으로 인한 글로벌 식량공급에 대한 우려가 높아지는 가운데 농업의 새로운 성장동력을 찾기 위한 방안으로 스마트팜(Smartfarm)에

* 정회원, 경희대학교 빅데이터응용학과
** 정회원, 경희대학교 글로벌경영학과
이 논문은 2024년 석사학위논문의 일부를 수정·보완한 것임.

접수일자 : 2025년 10월 31일
수정일자 : 2025년 11월 17일
게재확정일 : 2025년 11월 25일
교신저자 : 홍아름 e-mail : arhong@khu.ac.kr

대한 관심이 지속되고 있다. 스마트팜은 비닐하우스, 유리온실, 축사 등에 ICT, 빅데이터, 인공지능, 로봇 등 4차 산업혁명기술을 접목하여 작물과 가축의 생육환경을 원격으로 적정하게 유지·관리할 수 있는 농장[1]을 말한다. 스마트팜 시장 규모는 수요 증가에 따라 급속한 성장을 이루었다. 2022년 글로벌 스마트팜 시장 규모는 150.6억 달러 수준[2]이며, 한국의 스마트팜 시장 역시 2020년 2.4억 달러에서 2025년 4.9억 달러로 연평균 15.5% 이상 성장할 것으로 전망[1]하고 있다. 하지만 스마트팜 도입과 운영을 위해서는 각종 IoT(Internet of Things)장비 초기 구매 비용과 운영관리 능력이 필요하다. 스마트팜 IoT장비는 햇빛·비 등 외부환경에 노출되기 쉬우며 고장이 났을 경우 재빠른 조치를 취하지 않으면 농작물 재배에 치명적 피해를 받을 수 있다. 이러한 현실적 이유로 한국의 시설원예 스마트팜 농가들은 상대적으로 도입과 운영이 용이한 온도센서(내부 98.0%, 외부 83.6%), 습도센서(내부 96.0%, 외부 69.0%), CO_2센서(68.7%), 일사량센서(내부 54.0%, 외부 69.2%) 등 몇 가지 주요한 시설·장비를 중심[3]으로 농장을 운영하고 있다.

최근 시설원예 분야 스마트팜에 관한 연구, 특히 머신러닝과 딥러닝 알고리즘을 이용한 연구가 활발해지고 있지만, 다양한 실험적 연구가 농가에 직접적으로 활용되기 어려운 상황에 주목했다. 본 연구는 스마트팜 도입 농가의 현실적 조건을 고려함으로써 실제 농가에서 일반적으로 수집하고 활용할 수 있는 데이터와 적합한 분석 도구를 활용해 이를 해당 농가에 맞게 적용하는 연구를 제안한다.

본 연구의 목적은 첫째, 한국의 스마트팜 온실의 주요 시설·장비 데이터 변수를 활용해 토마토 생산량 예측을 위한 최적 환경 변수를 선정하여 범용성 있는 모델을 제안하는 것이다. 둘째, 농부 개인의 역량에 의존한 관행 농업의 한계를 극복하고 정밀 농업(Precision Agriculture)의 실현을 통해 적은 노동력과 자원을 투입하면서도 경험이 풍부한 베테랑 농업인과 비등할 정도의 생산성을 달성하는데 기여하는 것이다. 셋째, 토마토 생육에 가혹한 환경에서 재배된 데이터를 기반으로 기후 변화에 대비한 스마트팜 연구 데이터의 다양성을 확보하는 것이다.

Ⅱ. 본 론

1. 이론적 배경과 선행연구 검토

분석에 앞서 국내외 토마토 등 시설원예 스마트팜 작물의 생산량 예측과 관련한 선행연구를 검토했다. 홍성은 외(2020)[4]는 4개 토마토 농가의 환경 및 생육 데이터를 대상으로 하여, 다중선형회귀, 랜덤포레스트, 딥러닝(ConvLSTM) 알고리즘을 적용했다. 그 결과 LSTM 모델에 Convolution 레이어를 추가한 ConvLSTM 모델의 R^2 점수가 생산량 예측 0.981, 생장량 예측 0.805로 가장 높았다. 이세연 외(2023)[5]는 농촌진흥청 데이터를 대상으로 하여, LSTM, GRU, BI-LSTM 모델을 적용했다. 종속변수는 열매 수로 하였으며 독립변수는 개화화방, 내부온도, 줄기굵기, 엽수, 내부CO_2로 선정했다. 실험 결과 BI-LSTM 모델이 RMSE 1.952, MAPE가 0.082 높은 성능을 보였다. 나명환 외(2017)[6]는 2014년 연동 비닐온실에서 재배된 토마토를 분석 대상으로 하였다. 이 연구에서는 주요 환경 데이터를 독립변수로 하고, 단위 면적당(㎡) 일주일 누적 평균 수확량을 종속변수로 하여 다중회귀분석을 시행하였고, 시간적 지연 효과를 반영하여 토마토의 수확량을 수확 1주전의 평균온도, 7주전의 일사량, 5주전의 물 흡수량과 7주 전의 공급PH로 설명하는 모형을 완성했다. 강수람 외(2021)[7]는 55개 스마트팜 토마토 농가 데이터를 바탕으로 시계열 분석을 수행했다. 내부온도, 외부온도, 내부습도, CO_2농도의 주별 평균·최소·최대로 이루어진 외생변수와 주별 생산량을 변수로 삼았다. 분석에 사용된 모델은 LSTM, Input Attention LSTM, Dual Attention LSTM인데, 11주의 데이터를 학습하여 다음 1주의 생산량을 예측하는 Dual Attention LSTM 모델이 RMSE 1.01로 가장 우수한 예측력을 보였다. 김성란 외(2018)[8]는 2017년 토마토 데이터를 대상으로 품질 및 생육에 미치는 요인을 분석

하였다. 이때 검토된 변수는 내부온도, 외부온도, 일사량, 습도, 감우, CO_2농도, 양액EC와 PH 등 환경 데이터와 초장, 생장 길이, 엽수 엽폭 등 생육 데이터였다. 연구결과 내부온도, 누적일사량, 개화군, 수확군이 생산량에 영향을 미치는 주된 요인으로 나타났다. Alhnaity et al. (2019)[9]은 영국에 위치한 온실 데이터를 바탕으로 토마토 생산량 예측 모델을 제안했다. 이 연구에서 사용한 모델은 Support Vector Regression(SVR), 랜덤포레스트, LSTM이었다. 세 모델 중에서 LSTM 모델이 MSE 0.002, RMSE 0.047, MAE 0.03으로 가장 높은 성능을 나타냈다.

선행연구에서 활용한 머신러닝 및 딥러닝 기반 모델은 대체로 우수한 성능을 나타냈지만, 모델별로 사용한 독립변수와 종속변수의 대상과 수가 각기 달랐다. 어떤 기준으로 모델의 변수를 선정할 것인지 선정 기준에 대한 판단이 필요할 것으로 보였다. 또한, 선행연구에서는 생육 데이터 수집의 한계로 주별(week) 데이터를 활용한 연구가 많았다. 한편 한국을 대상으로 한 연구에서 활용된 데이터 세트를 보면 주로 따뜻한 지방에서 수집된 데이터 세트가 많았다. 토마토는 일조량이 좋고 온실 온도 관리가 용이한 지역에서 재배하기 유리하기 때문이다. 이상 선행연구를 검토한 결과를 바탕으로 기존 연구와 차별화된 본 연구의 방향과 모형을 설정했다.

2. 연구 모형

본 연구는 국내 스마트팜 온실에서 재배되는 토마토 환경·생육 데이터를 바탕으로 머신러닝·딥러닝 알고리즘을 사용하여 토마토 생산량 예측에 중요한 변수가 무엇인지 분석하는 것을 목적으로 한다. 이에 머신러닝 기법인 랜덤포레스트(RandomForest) 모델[10]과 RNN에 기반한 딥러닝 알고리즘인 LSTM(Long-Short Term Memory) 모델[11] 및 GRU(Gated Recurrent Unit) 모델[12]을 사용하였다. 데이터는 안정적인 품질과 시설·장비의 다양한 특성을 고려할 수 있도록 겨울철 한국 스마트팜 온실에서 수집되어 공개된 데이터를 대상으로 하였다.

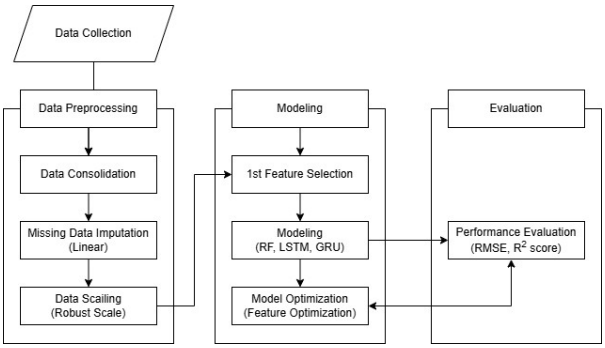

그림 1. 연구흐름도

3. 연구 내용

가. 데이터 선정 및 전처리

본 연구는 한국의 농림수산식품교육문화정보원이 운영하는 스마트팜 데이터 마트에 공개된 2023년 경진대회 데이터 세트를 대상으로 하였다. 해당 데이터는 다른 공공 데이터 세트에 비해 비교적 최신 자료이고, 상대적으로 변수의 수가 많으며 결측치는 적기 때문이었다. 또한 폭설이 빈번한 지역에서 겨울철에 수집된 데이터라는 점도 이번 데이터 세트를 선정한 요인으로 작용했다. 농업 분야 데이터는 크게 환경 데이터와 생육 데이터로 구분할 수 있는데, 주로 센서에서 수집되는 환경 데이터는 시간대별 데이터가 수집되는 데 반해, 생육 데이터는 조사원이 직접 조사를 해야 하므로 수집 간격이 고르지 못하고 조사 방법 및 조사원 특성에 따라 데이터의 품질 유지가 어려운 부분이 있다. 경진대회 데이터는 조사원 1명이 약 4개월간 같은 방법으로 약 5~20개의 표본을 정기적으로 수집한 데이터로서 다른 데이터 세트에 비해 장점이 있다고 판단했다.

해당 데이터 세트 중 환경 데이터는 내부온도, 내부습도, 내부CO_2, 토양온도, 토양습도, 누적일사량, 일사량, 외부온도, 외부습도, 풍향, 풍속, 강우신호, 외부일사량, 공급EC, 배액EC, 공급PH, 배액PH로 구성되었다. 이 중에서 수집된 데이터가 부족한 데이터는 항목에서 제외하였으며 생육 데이터는 데이

터 수집 기간이 평균 주 1회 전문재배사를 통해 수집된 정보를 사용하였다. 환경 데이터는 세팅에 따라 초 단위로 기록되었기 때문에 1주일마다 기록되는 생육 데이터와 단위를 통합할 필요가 있었다. 데이터 통합(Data Consolidation) 단계에서 환경 데이터와 생육 데이터를 일(day)별로 단위를 일치시켰다. 환경 데이터는 일별 평균(mean)치를 계산했으나 누적일사량(SRS)은 일별 합계(sum)로 집계하여 누적의 물리적 의미를 유지했다. 생육 데이터의 부족한 데이터는 선형 보간법(linear interpolation)을 통해 보완했다. 생육 데이터는 조사일에 5~20개의 샘플을 대상으로 조사된 데이터로 일별 평균값으로 처리하였다. 시계열 모델의 데이터는 시간적 인과관계를 명확히 하기 위해 라그(lag) 구조를 적용했다. 즉, t-6~t일의 환경 데이터로 t+1일의 생산량을 예측하도록 데이터를 재구성함으로써, 과거의 환경 제어 결과가 미래의 생산량에 미치는 영향을 정확히 모델링할 수 있도록 했다. 총 샘플 수는 123일 간의 데이터로 날짜 범위는 2023년 10월 13일부터 2024년 2월 13일이며, Robust Scale 방법을 통해 데이터 스케일링을 진행했다.

나. 변수 선정

준비된 데이터 중에서 이번 연구에 사용할 데이터의 변수를 선정하였다. 이번 연구에서의 독립변수는 환경 데이터, 종속변수는 생육 데이터 중에서 열매수(frtstCo)와 수확 수(hvstCo)의 합(outtrn)으로 정하였다. 선행연구에서는 토마토 생산량 예측을 위한 독립변수로 환경 데이터와 생육 데이터를 함께 반영하거나, 종속변수로 열매 수와 잎의 길이(lefLt)를 함께 적용한 연구[4]도 있었다.

하지만 주기적으로 생육 데이터를 수집하기 어려운 대다수 스마트팜의 현실을 생각했을 때, 생산량 예측 모델의 독립변수는 센서 데이터로 자동 수집되는 환경 데이터에 한정하는 것이 향후 모델의 활용 가능성 측면에서 유용하다고 판단했다. 종속변수로는 열매 수와 수확 수의 합산 값을 사용했다. 이는 농업인의 경영적 판단에 의한 수확량이 토마토 재배 데이터의 결과 값에서 배제되는 것을 피하기 위함이었다. 농업인의 판단에 따라 토마토의 숙성 정도에 따른 수확 수에 차이가 발생하기 마련인데, 환경 데이터에 의한 토마토 생산량을 전체 데이터 수집 기간을 통해 통합적으로 파악하기 위해선 아직 온실에서 재배되는 열매 수와 수확되어 출하된 수확 수, 두 항목을 합산하여 반영하는 것이 필요하다.

독립변수와 종속변수의 범위를 결정한 후에는 실제 구체적으로 모델을 학습할 세부 항목을 선정했다. 변수 선정의 기준은 한국 시설원예 스마트팜 시설·장비 설치 현황조사[3]에서 개별 농가의 설치 비율이 높은 장비로 하였다.

표 1. 입력 데이터의 특성

구분		변수명	코드명	장비설치현황
독립변수	실내외 환경 데이터	내부CO_2(ppm)	CI	68.7%
		내부습도(%)	HI	96.0%
		외부습도(%)	HO	69.0%
		공급EC(dS/m)	EI	66.5%
		배액EC(dS/m)	EO	49.6%
		공급PH(ph)	PI	61.7%
		배액PH(ph)	PO	45.6%
		누적일사량(MJ/㎡)	SRS	69.2%
		내부온도(℃)	TI	98.0%
		외부온도(℃)	TO	83.6%
종속변수	생산량	열매 수 + 수확 수(개)	outtrn	

위에서 선정된 열 가지 변수를 바탕으로 생산량 예측에 가장 최적화된 변수의 조합을 분석했다. 모든 가능한 변수의 조합(N)으로 모델의 성능을 평가했고 각 성능 지표를 저장했다. 그 결과 공집합을 제외한 모든 부분집합의 수는 1,023개였고, 3개의 모델에 각각 학습을 시켰을 때, 분석 대상은 3,069개의 결과 값이 도출된다.

다. 모델링

모델링을 위한 환경은 Google Colab에서 지원하는 라이브러리를 활용하였다. 웹에서 구동되어 가상머신을 통해 서버를 구축하는 Colab은 큰 비용적

부담 없이 농업 현장에서 활용하기에 적합하다. 본 연구에서 사용된 Colab 환경은 다음과 같다.

표 2. 서버환경 및 프로그램 환경

환경	구분	내용
서버 환경	GPU	T4GPU 15G
	Memory	51GB
프로그램 환경	Python	3.10.12
	TensorFlow	2.15.0
	Pandas	2.0.3

토마토 생산량 예측을 위해 사용되는 데이터는 특정 시간에 측정되거나 수집된 데이터가 대부분으로, 시간에 따른 환경 값의 변화와 작물의 성장 추이를 나타낸다. 따라서 본 연구에서도 다변량 시계열 예측 모델 중에서 가장 대표적인 LSTM 모델과 이보다는 단순한 형태인 GRU 모델을 사용하였다. 머신러닝 모델 중에서는 랜덤포레스트 모델을 활용하였다. 농업 분야 데이터에서 환경 데이터는 주로 센서 데이터인데, 센서 데이터는 종종 불균형한 클래스 분포를 가질 수 있다. 랜덤포레스트 모델은 여러 개의 결정 나무를 사용하는 기본구조[13]때문에 이러한 클래스 분포의 불균형성을 일부 보완할 수 있다. 본 연구에서 사용된 각 모델별 파라미터는 다음과 같다.

표 3. 모델별 파라미터

랜덤포레스트		LSTM		GRU	
n_Estimators	100	Units	50	Units	50
		Optimzer	Adam	Optimzer	Adam
Max_Depth	None	Epochs	50	Epochs	50
Max_Features	10	Learning_rate	0.01	Learning_rate	0.01
		Batch Size	32	Batch Size	32
MinSample_Split	2	Lookback(L)	7일	Lookback(L)	7일
MinSamples_Leaf –	1	Horizon(H)	1일	Horizon(H)	1일
		Stride	1일	Stride	1일

라. 성능 평가

1,023개의 변수 조합을 각 모델별로 학습한 생산량 예측 결과 중 가장 높은 성능을 달성한 모델별 변수는 다음과 같다.

표 4. 모델별 성능 비교

모델	사용변수	RMSE	R2
랜덤포레스트	CI, EI, SRS, TO	1.4575	0.8862
LSTM	EI. SRS, HO	3.6296	0.1024
GRU	PI, PO, HO	3.6593	0.0876

라그(lag) 구조를 적용한 결과, 랜덤포레스트 모델은 우수한 성능을 유지하며, 4개의 변수만으로도 높은 예측력을 보였다. 반면, RNN 계열 딥러닝 모델인 LSTM과 GRU는 라그 구조 적용 및 Lookback 설정으로 인해 학습 가능한 샘플 수가 감소하면서 예측 성능이 크게 낮게 나타났다. 이는 RNN 계열 모델이 효과적으로 학습하기 위해서는 충분한 시계열 데이터가 필요하며, 본 연구에서 사용된 약 4개월간의 데이터는 시퀀스 생성 과정에서 샘플 수가 감소하여 모델 학습에 제약이 있었음을 시사한다.

랜덤포레스트 모델에서 가장 높은 성능을 보인 변수 조합을 토대로 생산량 예측 값(Predict)과 실제 값(Actual)을 비교해서 나타낸 그래프는 다음과 같다.

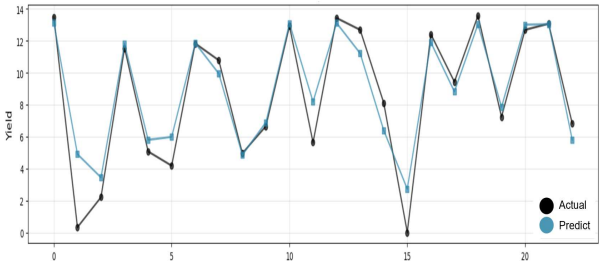

그림 2. 랜덤 포레스트 모델 최고 변수 조합의 예측값

3개 모델 중 유의미한 성능을 보인 랜덤포레스트 모델의 변수 조합에 따른 모델 성능은 다음과 같다.

표 5. 랜덤포레스트 모델 성능 상위 10개 변수 조합

순위	변수	RMSE	R2
1	CI, EI, SRS, TO	1.4575	0.8862
2	CI, EI, SRS, TI, TO	1.5171	0.8767
3	CI, EI, EO, HI, TI, TO	1.5222	0.8759
4	CI, EI, SRS, TI	1.5236	0.8757
5	CI, EO, TI	1.5255	0.8753
6	PI, SRS, TI, TO	1.5287	0.8748
7	CI, EI, EO, TO	1.5463	0.8719
8	CI, EI, HO, SRS, TO	1.5580	0.8700
9	CI, EI, EO, TI, TO	1.5647	0.8689
10	HO, PI, PO, SRS, TI, TO	1.5650	0.8688

랜덤포레스트 모델 중에서 변수별 성능이 가장 높은 조합은 내부CO_2, 공급EC, 누적일사량, 외부온도 네 가지 변수를 학습했을 경우였다. 위 네 가지 변수를 학습한 랜덤포레스트 모델의 토마토 생산량 예측은 RMSE 1.4575, R^2 점수 0.8862를 나타냈다. 하지만 그에 못지 않게 우수한 성능을 달성한 상위 10개 변수 조합도 있었다. 각 모델의 최적화에 필요한 경우의 수는 3개에서 6개까지고, 각 변수의 빈도수를 살펴보면 내부CO2(8회), 외부온도(8회), 내부온도(7회), 공급EC(7회), 누적일사량(6회)가 주요 변수로 도출되었다. 상위 10개 모델에서 전체 1,023개의 변수 조합 중 상위 10% 모델(102개)로 탐색 범위를 확장한 경우에도, 평균 변수 개수는 5.24개, 출현 빈도가 높은 변수는 위 5종으로 동일했다.

표 6. 성능 상위 10%(n=102)의 변수별 출현 빈도

모델	변수	출현	출현 비율
랜덤포레스트	TO	86	84.31%
	SRS	83	81.37%
	TI	73	71.57%
	CI	60	58.82%
	EI	56	54.90%
	PO	45	44.12%
	HO	44	43.14%
	HI	38	37.25%
	PI	34	33.33%
	EO	15	14.71%

Ⅲ. 결 론

본 연구 결과를 바탕으로 스마트팜 토마토 생산량 예측 모델의 최적 환경 변수 선정 연구에 대한 의미를 세 가지 제안으로 정리하면 다음과 같다.

첫째, 토마토 생산량 예측 모델에서 최적화 변수를 제안했다. 이번 연구의 차별적 의미는 토마토 생산량 예측 모델을 최적화하는데 일반적으로 예상하는 수보다 훨씬 적은 수의 변수가 필요하다는 점이다. 내부CO2, 공급EC, 외부온도, 내부온도, 누적일사량 변수는 본 연구에서 구성한 모델에서 주요 변수로 활용되었다. 반드시 많은 수의 변수가 최고의 모델을 보장하지 않는다는 결과는, 데이터 기반 최적화를 시도하는 농업인이 IoT장비 도입 과정에서 비용적·기술적 어려움에 직면했을 때, 어떤 장비를 우선적으로 도입할 것인가에 관한 의사결정을 지원해 줄 수 있다는 점에서 의의가 있다. 둘째, 약 4개월의 소규모 데이터셋에서는 LSTM, GRU와 같은 RNN 계열 모델보다 랜덤포레스트가 더 우수한 성능을 보였다. LSTM과 GRU가 유의미한 성능을 발휘하기 위해 충분한 시계열 데이터를 요구하는 반면, 농업 데이터는 품목, 작기, 수확시기 등의 제약으로 인해 장기간 데이터 확보가 어려울 수 있다. 따라서 제한된 데이터 환경에서는 랜덤포레스트가 보다 실용적인 선택임을 시사한다. 다만 향후 연구에서는 Mamba(2023)[14], Hawk & Griffin(2024)[15] 등 최신 시퀀스 모델링 아키텍처의 적용 가능성을 검토할 필요가 있다. 셋째, 본 연구는 토마토 생육에 불리한 기후조건에서 수집된 데이터 세트를 사용하여 사례의 다양성을 높였다. 이번 연구에서 예측에 활용된 데이터는 한국의 평창에서 수집되었는데, 평창은 최근 5년간 겨울철 평균 기온이 영하 4.9℃, 평균 적설량 88.1cm인 지역이다. 스마트팜은 비닐·유리로 차단된 환경을 구축하고 외부 날씨의 영향을 덜 받을 수 있는 장점이 있지만, 이러한 조건에서의 토마토 재배 전략은 일반적 농업경영 패턴과는 다를 수밖에 없다. 극단적 이상기후가 일상화될수록 가혹한 환경에서 생산량을 최적화하기 위한 노력이 중요하다는 점에서 본 연구 결과의 의미를 찾을 수 있을 것이다.

다만 본 연구는 경진대회에서 진행된 데이터로서

약 4개월간의 짧은 기간 동안 수집된 데이터로 장기적인 토마토의 생장을 살펴보기에는 다소 부족한 시간이었다는 점, 그리고 비용 데이터를 포함하지 않았다는 점에서 한계가 있다. 향후 토마토 도매시장 데이터 등을 고려하여 비용적 관점에서 토마토 생산량 최적화 모델을 구성하고, 초거대언어모델(LLM)을 활용하여 접근성과 편의성을 높인다면 농장을 경영하는 농업인들에게 더욱 직접적인 도움이 되는 모델이 될 수 있을 것이다.

REFERENCES

[1] 농림수산식품부 홈페이지(2025), https://www.mafra.go.kr/home/5280/subview.do, (accessed Oct., 23, 2025).

[2] Statista. Forecast market value of smart farming worldwide in 2021 to 2027. https://www.statista.com/statistics/720062/market-value-smart-agriculture-worldwide/?srsltid=AfmBOor57rXl0gNHdiMKhDWRvrTiw8mUVCDI94xjHGmxMYaEQIus8vwF, (accessed Oct., 23, 2025).

[3] 스마트팜코리아(2025), https://www.smartfarmkorea.net/board/list.do?menuId=null, (accessed Oct., 23, 2025).

[4] 홍성은, 박태주, 방준일, 김화종, “ConvLSTM 을 사용한 토마토 생산량 및 성장량 예측 모델에 관한 연구”, *한국정보기술학회논문지*, 제18권, 제1호, 1-10쪽, 2020년 1월

[5] 이세연, 양현정, 김민영, 김준경, 손아영, 홍성훈, “스마트팜 활용을 위한 BI-LSTM 기반의 토마토 생산량 예측에 관한 연구”, *한국통신학회논문지*, 제48권, 제4호, 457-468쪽, 2023년 4월

[6] 나명환, 박유하, 조완현, “스마트팜 데이터를 이용한 토마토 최적인자에 관한 연구”, *한국데이터정보과학회지*, 제28권, 제6호, 1427-1435쪽, 2017년 11월

[7] 강수람, 조완현, 나명환, “토마토 생산량 예측을 위한 Dual Attention LSTM”, *한국품질경영학회 추계학술발표논문집*, 제2021권, 58쪽, 2021년 10월

[8] 김성란, 박길석, 정수진, 최용조, 강동휘, 정지윤, ...정정석, “경남지역 토마토 생산량에 영향을 주는 품종별 요인 분석”. *한국원예학회 원예과학기술지*, 제36권, 별호, 93쪽, 2018년 10월

[9] B. Alhnaity, S. Pearson, G. Leontidis, S. Kollias, “Using deep learning to predict plant growth and yield in greenhouse environments”, *International Symposium on Advanced Technologies and Management for Innovative Greenhouses: GreenSys2019 1296*, pp. 425-432, 2019.

[10] L. Breiman, “Random forests”, *Machine learning*, vol. 45, no.1 , pp. 5-32, 2001.

[11] S. Hochreiter, J. Schmidhuber, “Long short-term memory”, *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[12] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling”, arXiv preprint arXiv:1412.3555, 2014.

[13] K. J. Archer, R. V. Kimes, “Empirical characterization of random forest variable importance measures”. *Computational statistics & data analysis*, vol. 52, no. 4, pp. 2249-2260, 2008.

[14] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *in Proc. First Conf. Language Modeling,* May 2024.

[15] S. De, S. L. Smith, A. Fernando, A. Botev, G. Cristian-Muraru, A. Gu, R. Haroun, L. Pascanu, J. Rae, A. Razavi, L. Weidinger, M. White, and C. Gulcehre, "Griffin: Mixing gated linear recurrences with local attention for efficient language models," arXiv preprint arXiv:2402.19427, 2024.

저 자 소 개

김진만(정회원)

2001년 서울대학교 사범대학 학사 졸업.

2024년 경희대학교 AI기술경영학과 석사 졸업.

현재 경희대학교 빅데이터응용학과 박사과정 재학.

<주관심분야 : 스마트팜, 빅데이터, 생성형AI>

홍아름(정회원)

2001년 서울대학교 기술경영경제정책대학원 석박사 졸업.

현재 경희대학교 테크노경영대학원 글로벌경영학과 부교수

<주관심분야 : 스마트기술, 산업정책, AI기술경영>