Regular paper | Journal of the KIECS. pp. 1403-1412, vol. 19, no. 6, Dec. 31. 2024, t. 128, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.6.1403

토마토 당도 예측을 위한 회귀분석과 시계열 분석 비교 연구

백남곤* · 김진성** · 최은성*** · 심춘보** · 정세훈***

Comparative Study of Regression and Time Series Analysis for Predicting Tomato Sweetness

Nam-Gon Baek* · Jin-Seong Kim** · Eun-Sung Choi*** · Chun-Bo Sim** · Se-Hoon Jung***

요 약

농산물의 품질 향상과 재배 과정 최적화를 위해 정확한 당도 예측이 중요하다. 현재 과일의 당도는 파괴적 측정 방식이 주로 사용되며, 비파괴당도계는 부분 측정과 표면 손상의 한계가 있다. 본 연구에서는 AI Hub의 지능형 스마트팜 데이터세트를 활용하여 토마토의 당도를 예측하기 위해 XGBoost, LightGBM을 이용한 회귀분석과 LSTM을 이용한 시계열 분석을 수행하였다. 실험 결과, 회귀분석 모델들은 모두 음수의 R^2 값과 큰 오차율을 보였으나, LSTM을 활용한 시계열 분석에서는 토마토 당도의 일반적 범위 대비 허용 가능한 수준의 예측 오차 MAE 0.224를 달성하였다. 이는 토마토의 당도가 생육 전반에 걸친 시계열적 특성과 밀접한 관련이 있음을 시사한다. 본 연구는 LSTM 모델을 통해 비파괴적 방식의 실용적 당도 예측이 가능함을 입증하였으며, 토마토 당도 예측에 있어 시계열 분석 기법의 효용성을 검증했다는 점에서 의의가 있다.

ABSTRACT

Accurate sweetness prediction has become increasingly important for improving agricultural product quality and optimizing cultivation processes. Currently, fruit sweetness is primarily measured through destructive methods, while non-destructive sweetness meters are limited by partial measurement and surface damage. This study performed regression and time series analyses to predict tomato sweetness using intelligent smart farm datasets from AI Hub. We conducted regression analysis using XGBoost and LightGBM, and time series analysis using LSTM. In the experimental results, regression models recorded negative R^2 values with high error rates. In contrast, time series analysis using LSTM achieved an acceptable prediction error of MAE 0.224 compared to the typical range of tomato sweetness. This suggests that tomato sweetness is closely related to time series characteristics throughout the growth period. This study demonstrates the feasibility of practical non-destructive sweetness prediction using the LSTM model and is significant in validating the effectiveness of time series analysis techniques for predicting tomato sweetness.

키워드

Tomato Sweetness Prediction, Regression Analysis, Time Series Analysis, Smart Farm
토마토 당도 예측, 회귀분석, 시계열 분석, 스마트팜

* 국립순천대학교 스마트농업전공(bng0611@nate.com)
** 극립순천대학교 IT바이오융합시스템전공
(1235037@s.scnu.ac.kr, cbsim@scnu.ac.kr)
*** 극립순천대학교 컴퓨터공학과(20162077@s.scnu.ac.kr)
*** 교신저자 : 국립순천대학교 컴퓨터공학과
· 접 수 일 : 2024. 10. 30
· 수정완료일 : 2024. 11. 20
· 게재확정일 : 2024. 12. 12
· Received : Oct. 30, 2024, Revised : Nov. 20, 2024, Accepted : Dec. 12, 2024
· Corresponding Author : Se-Hoon, Jung
Dept. Computer Engineering, Sunchon National University
Email : shjung@scnu.ac.kr

Ⅰ. 서 론

정확한 당도 예측은 농산물 품질을 향상시키고 재배 과정의 최적화를 위해 필수적이다. 최근 몇 년간 인공지능, 특히 딥러닝의 응용이 높은 정확도와 복잡한 데이터 처리 능력으로 인해 농업 분야에서 주목받고 있다[1]. 이는 국내 소비자들이 과일 구매 시 양보다는 질, 특히 당도와 숙성도를 우선시하는 경향과 맞물려 있다[2]. 이러한 소비자들의 선호 변화는 농민들이 과일 수확 기준을 바꾸도록 영향을 미쳤다.

과일의 품질을 결정하는 중요한 내적 특성에는 맛, 향기, 식감이 있지만, 특히 당도는 소비자들이 가장 중요하게 여기는 요소이다. 과일의 단맛을 평가하는 지표로 흔히 당도(Brix)가 사용되며, 이는 100g의 수용액에 포함된 가용성 고형물의 비율을 %로 나타낸 것이다[3]. 기존에는 관능 평가를 통해 단맛을 주관적으로 평가했으나, 현재는 과육을 착즙하여 굴절식 당도계로 당도를 측정하는 방식이 주로 사용된다. 굴절식 당도계는 정량적이고 객관적인 값을 제공하지만, 과육을 파괴하는 측정 방식에는 한계가 있다.

이를 해결하기 위해 비파괴당도계가 도입되었지만, 이 역시 일부 제한 사항이 존재한다[4]. 비파괴당도계는 빛 투과 방식을 사용해 측정하며, 근적외선 파장을 이용한다. 그러나 이는 특정 부위의 당도만을 측정하기 때문에 과일 전체의 평균 당도를 반영하지 못하고, 과일 표면에 손상이 발생할 수 있다.

본 연구는 딥러닝 모델을 활용하여 온실에서 재배되는 작물의 당도를 예측하는 모델을 개발하는 것을 목표로 한다. 본 연구에서는 과육을 파괴하지 않고 생육 과정에서 수집된 시계열 데이터를 딥러닝 모델로 분석하여, 과일의 당도를 예측하는 방안을 제시하고자 한다.

Ⅱ. 관 련 연 구

본 연구는 딥러닝 모델을 활용하여 온실에서 재배되는 작물의 당도를 예측하는 재배 과정의 최적화를 위해 크게 작물 데이터 분석 및 LSTM(Long Short-Term Memory)기반의 데이터 분석으로 구분하여 선행 연구를 분석한다.

2.1 머신러닝을 이용한 작물 데이터 분석

[5]의 연구에서는 스마트팜에서 수집 가능한 환경 및 생육 데이터를 중심으로 생육 데이터 간의 상관관계를 분석하고 토마토 작물의 수확량에 가장 큰 영향을 미치는 대표 변수를 규명하였다. 토마토 작물에 대한 작기별 데이터이며, 환경 정보로는 온도, 습도, 일사량, 풍량, 풍속 및 토양 정보 등이며, 생육 정보로는 생장길이, 줄기직경, 엽장, 엽폭, 엽수, 개화군, 착과군 및 열매 수 정보가 제공된다. 이러한 환경정보와 생육정보를 이용하여 다중회귀분석을 활용하여 생육 변수가 간 상관관계에 대하여 분석하였다. 본 연구를 위한 주요 생육정보는 생장길이, 줄기직경, 엽장, 엽폭, 엽수, 개화군, 착과군, 및 열매수로 정의하였다. 토마토 생육 데이터는 다중회귀분석 모델을 적용했을 경우 엽장과 엽폭의 VIF(Variance Inflation Factor) 값이 10보다 큰 값을 나타내어 심각한 다중공선성 문제가 존재하였다. 따라서 다중공선성이 가장 높은 엽장을 제거 후 다시 다중회귀분석 모델을 적용하여 다중회귀분석 모델로부터 추정된 결정계수의 변화가 크지 않은 결과를 기반으로 토마토 수확량과 생육 변수의 연관성 정도를 분석하였다.

[6]의 연구는 빅데이터와 IoT 기술을 적용하여 작물의 환경을 조사하고 데이터를 측정 및 가공하여 생산량의 증대와 작물의 품질 향상 방안을 제안하였다. 스마트팜 딸기 농장에서 수집된 환경 및 생산량 데이터를 분석하여 양액 흡수량-환경 데이터와 생산량-환경 데이터 간의 상관분석을 통해 연관성이 높은 환경 데이터로 생산량 예측하였다. 예측 방법을 릿지 회귀(Ridge regression), LightGBM, XGBoost를 사용하여 작물 생산량 예측 모델을 분석하였으며, 생산량과 연관성이 있는 환경 데이터를 상관 분석한 결과, 9가지 데이터 중 EC, 습도, 일조량, 온도 생육 정보가 가장 중요한 요인임을 알 수 있었다. EC를 제외한 나머지 데이터들의 상관계수가 크지 않았다.

[7]의 연구에서는 선형 보간법으로 처리할 수 없는 결측치의 한계점을 언급하였으며, 이러한 한계를 극복하기 위해 편차 평균 보간법을 제안하였다. 또한 생육 데이터는 주마다 수기로 측정하기 때문에 환경 데이터와 수집 방식에 차이가 있어 시간 단위가 일치하지 않는 문제점이 있어 데이터의 통합 과정에서 시계열적 특성을 반영하기 위해 다양한 변수를 유형별로 분

류하여 최적의 데이터세트 구축방법을 제안했다. 데이터 변수별 결측치 비율도 함께 고려하여 외부 온도, 외부 일사량, 내부 온도, 내부 습도, 잔존 CO2, 토양 온도를 포함하는 환경 데이터, 그리고 농가별로 개체 번호, 초장, 엽수, 엽장, 엽폭, 엽병장, 관부직경, 화방별착과수를 포함하는 생육 데이터를 사용했다. 시계열 데이터 변수 추출 시, 통계 개념과 작물의 특성을 기반으로 이동 창 통계량 유형과 주간 및 야간 유형을 만들어 최종 데이터세트를 구축하였다. 결측치 처리 및 변수 추출을 통해 다양한 데이터세트에 따른 모델을 구축하였으며, 이동 창 통계량 유형의 GRU(: Gated Reccurent Unit) 모델이 가장 우수한 성능을 나타냈다. GRU는 LSTM을 단순화한 구조로 update gate와 reset gate를 통해 시퀀스 종보를 효율적으로 처리하여 더 간단한 구조와 적은 파라미터를 가진다. 시계열 데이터의 단위를 변환할 때 원 데이터의 특성을 반영하고, 다양한 농가와 개체를 한 모델에 학습시키며 데이터 부족 문제를 보완한다는 점에서 의의가 있다.

[8]의 연구에서는 참외 모니터링을 위해 스냅샷 방식 초분광 영상의 적용 가능성을 검토하고, 초분광 데이터를 이용하여 참외의 중요한 품질 요인인 당도를 예측하는 연구를 제안하였다. 참외의 당도 예측을 위해 SVR(: Support Vector Regression)과 KR(Kernel-Ridge Regression)의 두 가지 모델을 이용하였으며, 두 가지 모델 모두 RBF (Radial Basis Function) 커널을 적용하였다. 학습 데이터의 학습 중 과적합 등의 문제를 방지하기 위해 본 연구에서는 5겹 교차검증(5-fold cross validation) 방법을 적용하여 예측 모델의 튜닝 및 검증을 수행하였으며, 이를 통해 개발된 모델은 검증 데이터를 이용해 모델의 R^2(Coefficient of determination)과 MSE(: Mean Squared Error)로 평가하였다. SVR과 KR모델의 최적의 하이퍼파라미터를 결정하기 위하여 Gridsearch 방법을 적용하였다. 그 결과, SVR의 최적의 C와 gamma는 각각 1.0, 9.0이었으며, KR의 최적의 alpha와 gamma는 각각 0.55, 0.1이었다. 튜닝 된 SVR과 KR 모델을 이용하여 참외의 당도 값을 예측하였으며, SVR 모델의 R^2와 MSE는 각각 0.86, 0.85로 비교적 높은 정확도에서 참외의 당도 예측이 가능한 것으로 나타나 참외 당도 예측을 위해 스냅샷 방식 초분광 영상이 이용될 수 있음을 확인하였다.

2.2 LSTM기반의 데이터 분석

LSTM은 순차적으로 데이터 처리를 위해 이전 시점의 정보를 현재 계산에 반영하는 RNN(Recurrent Neural Network)의 한 종류로, 긴 시간 간격의 의존성을 처리하기 위해 고안된 구조입니다. 주요 구성 요소로는 셀 상태(Cell State)와 입력 게이트(Input Gate), 포겟 게이트(Forget Gate), 출력 게이트(Output Gate)가 있다. 셀 상태는 정보를 장기적으로 저장하며, 입력 게이트는 새로운 정보를 얼마나 반영할지 결정하고, 포겟 게이트는 불필요한 정보를 제거한다. 출력 게이트는 최종 출력을 생성한다. LSTM은 텍스트, 음성, 시계열 데이터와 같은 순차적 데이터를 효과적으로 처리할 수 있다[9].

[9]의 연구에서는 스마트팜에서 토마토 생산량을 예측하기 위한 방법론을 개발하고, 이를 실사용할 수 있또록 웹 서비스 구현을 제안하였다. [6]의 연구에서는 스마트팜 내부 환경에 집중하여 외부 온도 등 외부 환경 요소는 제외하고, 종속 변수로 열매 수를 제안했다. 상관관계 분석 결과, 내부 온도, CO2, 누적 일사량, 생장 길이, 엽수, 엽폭, 줄기 굵기, 개화화방 등이 열매 수와 유의미한 상관관계를 제시하였다. 예측 모델로는 다중 선형 회귀, 의사결정 나무(Decision Tree Regression), 랜덤 포레스트(RandomForest), LightGBM, XGBoost 등을 사용하였다. 실험 결과, BI-LSTM 모델이 가장 우수한 성능을 보였으며, RMSE(: Root Mean Squared Error)가 1.952, MAPE(: Mean Absolute Percentage Error)가 0.082로 다른 알고리즘보다 낮았다[11].

[10]의 연구에서는 LSTM 네트워크를 활용하여 도시 지역의 대기 오염을 비롯한 다양한 환경 요소들을 예측하는 방법을 제안하였다. 데이터 전처리 과정에서 원시 데이터를 정제하고 변수들의 스케일을 조정하여 학습 효율을 높였으며, 바람 방향과 날짜 데이터는 분석에 맞게 변환되었다. LSTM 모델은 RMSE 0.06680을 기록하며, 선형 회귀 모델(RMSE 0.07022)보다 더 낮은 오차를 보였다. 이는 LSTM이 대기 오염 시계열 데이터의 복잡한 패턴을 효과적으로 모델링할 수 있다는 연구 결과를 제시하였다.

표 1. 스마트 팜 센서 및 토마토 생육 데이터
Table 1. Smart farm sensors and tomato growth data

Information	Type of information
Sensor data	Absolute humidity 1, Absolute humidity 2, CO_2 concentration, Water vapor density in air, humidity percentage, Celsius temperature, Accumulated radiation, Instantaneous radiation, Supply pH, Supply amount, Drainage EC, Drainage pH, Drainage amount, Number of supplies, Amount per supply, Drainage rate.
Growth data	Fruit width, Fruit height, Fruit weight, Sugar content, Acidity, Firmness, moisture content.
Number of data	600

[12]의 연구에서는 기후 변화에 따른 농업 환경의 변동성을 고려하여, 신뢰할 수 있는 데이터 분석 모델을 개발하고, 이를 통해 농업 생산의 안정성을 높이는 연구를 제안하였다. 토양수분, 온도, 습도 등의 환경 요인은 식물의 성장과 수확량에 직접적인 영향을 미치며, 이러한 요인들의 정확한 분석과 관리는 농업의 지속 가능성을 보장하는 데 필수적이다. 토양수분을 중심으로 온도와 습도 등 다른 환경 변수들을 종합적으로 분석하며, 해당 데이터를 기반으로 농작물 관리 전략을 최적화하는 방안을 제안하였다. 토양수분을 중심으로 온도와 습도 등 다른 환경 변수들을 종합적으로 분석하며, 해당 데이터를 기반으로 농작물 관리 전략을 최적화하는 방안을 제안하였다. 특히, 기후 변화에 따른 농업 환경의 변동성을 고려하여, 신뢰할 수 있는 데이터 분석 모델을 제안하였다. [8]의 연구에서 제안된 LSTM 모델의 성능으로 RMSE는 4.2706으로 측정되었으며, MAE(: Mean Absolute Error)는 3.089로 측정되었다. 모델의 예측 성능을 심층적으로 분석하기 위해 잔차의 분포를 조사하였으며, 잔차들이 정규 분포에 가깝게 나타났음을 확인하였다.

III. 당도 예측 모델

당도는 농가의 수익에 직결되는 핵심 요소이므로, 센서 및 생육 데이터를 통해 당도를 예측할 수 있다면 농가의 수익성 증대를 도모할 수 있다. 본 장에서는 당도 예측을 위해 데이터세트를 훈련에 적합한 형태로 전처리한 후, 회귀분석과 시계열 분석을 수행한다. 토마토 당도 예측 모델 프로세스는 그림 1과 같다.

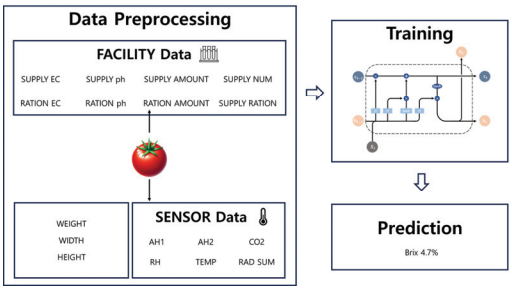

그림 1. 토마토 당도 예측 프로세스
Fig. 1 Tomato sweetness (brix) prediction process

3.1 데이터 세트 및 전처리

본 연구에서는 AI Hub에서 제공하는 지능형 스마트팜 토마토 통합 데이터를 활용하였다[13]. 해당 데이터는 이미지 데이터와 시계열 데이터가 통합된 형태로 제공되며, 본 연구에서는 스마트팜 내부의 센서 데이터와 토마토 생육 데이터를 주요 분석 대상으로 선정하였다. 원본 데이터에는 다수의 농가 정보가 포함되어 있으나, 각 농가별 특성 차이와 결측치 문제를 고려하여 결측치가 가장 적은 1개 농가의 데이터만을 선별하여 활용했다. 센서 데이터와 생육 데이터는 스마트팜 내부의 환경을 모니터링 하는 다양한 센서들로부터 수집된 정보와 생육과 관련한 데이터를 포함하고 있으며, 구성은 표 1과 같다. 센서 데이터는 온도, 습도, CO2 농도, 일사량 등 작물 생육에 직접적인 영향을 미친 환경 요인들을 측정한 값들로 구성되어 있다. 생육 데이터의 경우, 전체 농가에서 무작위로 선별된 100개의 토마토 표본에 대해 파괴 검사를 통해 수집된 과폭, 과고, 과중, 당도, 산도, 경도, 수분율이 포함돼있다. 이러한 생육 데이터

는 모든 농가의 특성을 대표할 수 있는 표본으로 간주하여, 선별된 1개 농가의 센서 데이터와 통합하는 과정을 수행했다. 통합된 데이터의 개수는 600개다.

데이터 전처리는 크게 두 단계로 진행되었다. 첫 번째 단계는 데이터 통합 과정이다. 5분 단위로 기록된 원본 센서 데이터와 1주차 단위로 기록된 생육 데이터의 시간 단위를 일치시키기 위해, 센서 데이터를 주차별 평균값으로 변환하여 통합하였다. 이후 시간을 기준으로 두 데이터를 통합하여 분석용 데이터세트를 구성했다. 두 번째 단계는 상관관계 분석을 통한 특성 선별 과정이다. 통합된 데이터세트에서 당도 예측에 유의미한 영향을 미치는 변수들을 선별하기 위해 피어슨 상관계수를 활용하였다. 그림 2는 상관관계 분석 결과다.

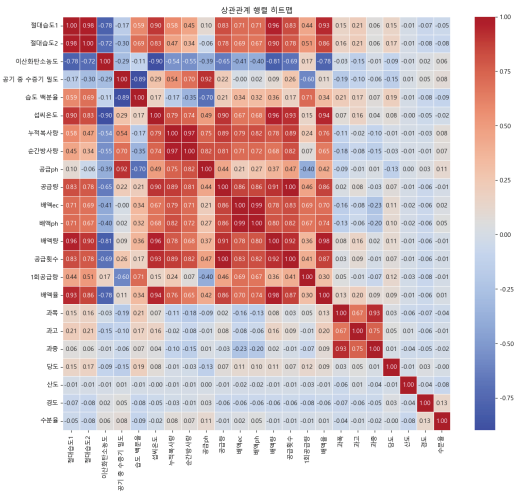

그림 2. 상관관계 분석 결과
Fig. 2 Correlation analysis results

최종적인 분석결과에 따르면 각 특성과 당도의 상관관계 현저히 낮게 측정되었다. 이는 당도가 단일 변수와의 단순 선형 관계가 아니라 환경요인과 생육 특성 등 복잡한 상호작용에 의해 결정될 수 있음을 시사한다. 이러한 한계에도 불구하고, 상대적으로 상관관계가 높은 변수와 도메인 지식을 바탕으로 변수를 선별하였다. 센서 데이터에서는 습도 백분율(0.19), 절대습도1(0.15), 절대습도2(0.17), 섭씨온도(0.08), 배액량(0.11), 배액EC(0.11), 배액pH(0.10), 공급pH(-0.13)를 선정하였고, 생육 데이터에서는 과고(0.05), 과폭(0.03), 과중(0.01)을 분석에 포함하였다. 단, 이러한 낮은 상관관계는 시계열 데이터의 특성을 고려할 때 단순 상관관계 분석만으로는 변수 간의 관계를 완전히 설명하기 어려울 수 있음을 시사한다. 이러한 변수들을 종합하여 회귀 분석용 데이터세트를 구성하였다. 한편, 시계열 분석에서는 단순 상관관계 분석을 통한 컬럼 선별 작업이 의미가 없기 때문에, 전체 변수를 포함하여 통합한 데이터세트를 활용하였다.

3.2 당도 예측 모델

본 연구는 토마토의 당도 예측을 위해 회귀분석과 시계열 분석을 병행하여 수행하였다. 회귀분석에는 그래디언트 부스팅 기법 기반의 XGBoost와 LightGBM을, 시계열 분석에는 LSTM을 활용하였다. 그림 3은 딥러닝 기반 당도 예측 모델 구조다. 회귀분석과 시계열 분석을 별도로 진행한다.

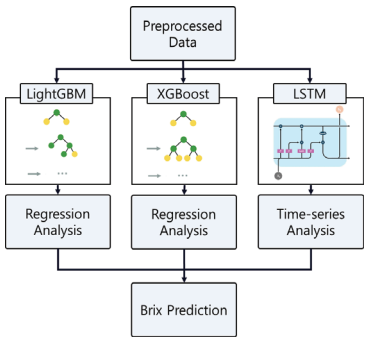

그림 3. 딥러닝 기반 당도 예측 모델 구조
Fig. 3 Deep learning-based sweetness (brix) prediction model architecture

회귀분석을 위해 활용된 XGBoost와 LightGBM은 트리 기반의 앙상블 학습 방법으로, 각각 여러 약한 학습기를 순차적으로 생성하여 강한 학습기를 만드는 부스팅 알고리즘이다. XGBoost는 그래디언트 부스팅의 속도와 성능을 개선한 모델로, 과적합 방지를 위해 정규화 항을 포함하고 있다. 반면, LightGBM은 리프 중심의 트리 분할 방식을 사용하여 XGBoost보다 학습 속도가 빠른 장점이 있다.

시계열 분석을 위한 LSTM은 순환 신경망의 한

종류로, 장기 의존성 문제를 해결하기 위해 설계된 모델이다. LSTM은 입력 게이트, 망각 게이트, 출력 게이트를 통해 시계열 데이터의 장기적인 패턴을 효과적으로 학습할 수 있다. 본 연구에서는 두 개의 LSTM 층을 사용하여 모델을 구성하였다. 첫 번째 LSTM 층에서는 Return sequences를 True로 설정하여 시퀀스 정보를 유지하도록 하였고, 두 번째 LSTM 층에서는 최종 예측값 생성을 위해 Return sequences를 False로 설정하였다. 과적합 방지를 위해 Dropout 층을 추가하였으며, 하이퍼파라미터 탐색 범위는 LSTM 유닛 수를 32, 64, 128로 설정하였고, 옵티마이저는 Adam과 RMSprop을 비교하였다. 손실 함수로는 MSE를 사용하였다.

모든 모델의 성능 평가는 예측값과 실제 값의 차이의 제곱 평균인 MSE, MSE의 제곱근으로 실제 단위의 오차를 측정하는 RMSE, 예측값과 실제값의 절대 차이의 평균인 MAE, 모델이 설명하는 분산의 비율인 R^2를 통해 수행되었으며, 회귀분석 모델의 경우 전체 특성을 사용한 경우와 상관관계 분석을 통해 선별된 특성만을 사용한 경우를 모두 평가하였다. LSTM 모델의 경우 시계열 데이터의 특성을 고려하여 전체 변수를 포함한 데이터세트를 활용하였다.

Ⅳ. 실험 결과

본 연구에서는 토마토의 당도를 예측하기 위해 XGBoost와 LightGBM을 이용한 회귀분석과 LSTM을 활용한 시계열 분석을 병행하여 수행하였다.

표 2는 회귀분석 결과다. 회귀분석에서는 XGBoost와 LightGBM 모델을 사용하여 실험을 진행하였다. XGBoost와 LightGBM 모델은 모두 GridSearchCV를 활용하여 최적의 하이퍼파라미터를 탐색하였다. XGBoost 모델의 경우, Colsample bytree는 1.0, Gamma는 5, Max depth는 3, Min child weight는 5, Subsample은 1.0으로 설정되었다. LightGBM 모델은 Colsample bytree를 1.0, Max depth를 10, Min child weight를 1, Subsample을 0.6으로 설정하여 모델의 일반화 성능을 향상시켰다. 실험 방식은 모든 특성을 사용한 경우와 상관관계 분석을 통해 선별된 특성만을 사용한 경우를 비교하였다.

표 2. 회귀분석 결과
Table 2. Regression Analysis Results

Model	MSE	RMSE	MAE	R^2
XGBoost (all columns)	0.638	0.799	0.391	-2.075
LightGBM (all columns)	0.339	0.582	0.346	-0.633
XGBoost (selected columns)	0.638	0.799	0.391	-2.075
LightGBM (selected columns)	0.352	0.593	0.369	-0.694

전체 특성을 사용한 경우, XGBoost 모델은 RMSE 0.799, MAE 0.391, MSE 0.638, R^2 -2.075를 기록했으며, LightGBM 모델은 RMSE 0.582, MAE 0.346, MSE 0.339, R^2 -0.633을 기록하였다. 상관관계 분석을 통해 선별된 특성만을 사용한 경우, XGBoost의 성능은 변화가 없었으며, LightGBM은 RMSE 0.593, MAE 0.369, MSE 0.351, R^2 -0.694를 기록하여 오히려 성능이 소폭 하락하였다. 두 실험 모두에서 음수의 R^2 값이 나타난 것은 모델이 평균값을 사용하는 것보다도 낮은 예측 성능을 보였음을 의미한다. 이러한 저조한 성능은 다음과 같은 특성에서 기인한 것으로 분석된다.

첫째, 토마토의 당도는 단순히 현재 시점의 환경 요인이나 생육 특성만으로는 설명하기 어려운 복잡한 특성을 가진다. 이는 모든 변수와의 상관계수가 0.2 미만으로 나타난 상관관계 분석 결과와도 일치한다. 둘째, 당도 형성은 생육 전 기간에 걸친 누적된 환경 요인의 영향을 받는다. 따라서 특정 시점의 데이터만을 고려하는 회귀분석으로는 이러한 시간적 종속성을 포착하기 어렵다.

특히 특성 선별 후 LightGBM의 성능이 오히려 소폭 하락한 것은, 낮은 상관관계를 보이는 변수들이라도 당도 예측에 있어 복합적인 영향을 미칠 수 있음을 시사한다. 이는 삭제된 습도, 온도, EC, pH 등의 변수들이 실제로는 당도 예측에 있어 복합적인 영향을 미칠 수 있음을 시사한다. 이러한 결과는 토

마토 당도 예측에 있어 단순한 회귀분석보다는 시계열적 특성을 고려할 수 있는 분석 방법이 더 적합할 수 있음을 시사한다. 실제로 이어지는 LSTM을 활용한 시계열 분석에서 이를 확인할 수 있다.

표3은 LSTM을 사용한 시계열 분석 결과다. 분석 결과 해당 데이터는 회귀 분석이 비해 시계열 분석이 적절한 것으로 사료된다. 시계열 분석을 위한 LSTM 모델의 최적 하이퍼파라미터는 GridSearchCV를 통해 Batch size는 32, Dropout rate는 0.2, Epochs는 100, Time steps는 1 Hidden units는 128로 설정되었다.

표 3. 시계열 분석 결과
Table 3. Time Series Analysis Results

Model	MSE	RMSE	MAE	R^2
LSTM	0.086	0.292	0.224	0.588

LSTM 모델의 실험 결과, MSE 0.086, RMSE 0.292, MAE 0.224를 기록했으며, R^2값은 0.588을 달성하였다. 이는 앞서 진행한 회귀분석의 음수 R^2 값과 비교할 때 통계적으로 유의미한 성능 향상을 보여준다. 성능 지표를 분석해보면, MSE 0.086과 이를 제곱근한 RMSE 0.292는 예측값과 실제 당도 간의 평균적인 오차를 나타내며, 이는 토마토 당도 예측에서 허용 가능한 수준의 오차 범위 내에 있다. MSE가 0.1 미만이라는 것은 예측값의 분산이 작아 모델이 안정적인 예측을 수행하고 있음을 시사한다. RMSE가 MAE보다 높게 나타난 현상은 예측값과 실제값 간의 큰 차이를 보이는 이상치가 존재함을 의미하며, 이는 생육 과정에서의 급격한 환경 변화나 특수 재배 조건에 따른 영향으로 추정된다. MAE 0.224는 예측값이 실제 당도와 평균적으로 0.22 Brix의 절대 차이가 보임을 의미하며, 이는 토마토 당도의 일반적 범위인 3~5 Brix를 기준으로 할 때 약 5.5%의 상대 오차율에 해당한다.

회귀분석 결과에서 입증된 바와 같이, 토마토의 당도는 단일 시점의 환경 요인이나 생육 특성만으로는 설명하기 어려운 복잡한 특성을 가진다. LSTM 모델이 회귀분석 대비 우수한 성능을 보인 것은 시계열 데이터의 시간적 의존성을 효과적으로 학습할 수 있는 모델 구조에 기인하나, R^2값 0.588이 나타내는 중간 수준의 설명력은 다음과 같은 기술적 한계에서 비롯된 것으로 분석된다.

첫째, 데이터의 시간적 해상도 불일치이다. 센서 데이터는 5분 단위로 수집되어 주차별로 평균화되었고, 생육 데이터는 1주차 단위로 기록되어 있다. 이러한 시간 단위의 불일치로 인한 정보 손실이 예측 정확도를 저하시키는 요인으로 작용하였다. 둘째, 시간적 종속성의 복잡성이다. 토마토의 생육 과정에서 이전 시점의 환경 요인들이 현재 시점의 당도에 미치는 영향이 비선형적 패턴을 보이며, 이러한 장기적 의존성을 모델이 완전히 포착하지 못한 것으로 분석된다. LSTM이 장기 의존성 학습에 최적화된 구조를 가지고 있으나, 데이터의 주차별 평균화 과정에서 발생한 정보 손실로 인해 모델의 성능이 제한된 것으로 분석된다. 셋째, 시계열 패턴의 비정규성이다. 토마토의 생육 과정에서 발생하는 환경 요인들의 시간적 변화가 불규칙적이고 비선형적인 특성을 보이며, 이는 예측의 복잡도를 증가시키는 요인으로 작용한다. LSTM 모델이 회귀분석과 달리 양의 설명력을 보인 것은 이러한 비선형적 시계열 패턴을 부분적으로 포착할 수 있었기 때문으로 해석된다.

본 연구 결과는 토마토 당도 예측에 있어 다음과 같은 의의와 시사점을 제공한다. 첫째, LSTM 모델이 달성한 5.5%의 상대 오차율과 0.588의 R^2 값은 비파괴 방식으로도 실용적 수준의 당도 예측이 가능함을 실증적으로 입증하였다. 이는 기존 파괴검사 방식의 한계를 극복할 수 있는 대안적 방법론을 제시한다. 둘째, 센서 데이터와 생육 데이터의 시간적 해상도를 일치시키고 수집 주기를 최적화함으로써 예측 성능의 향상 가능성을 확인하였다. 셋째, 당도 형성에 영향을 미치는 추가적 환경 요인 및 생육 특성 변수의 도입과 시계열 데이터의 전처리 고도화를 통해 모델의 설명력을 제고할 수 있는 구체적인 개선 방향을 제시하였다.

Ⅴ. 결론

본 연구에서는 토마토의 당도를 예측하기 위해 AI Hub에서 제공하는 지능형 스마트팜 데이터세트를 활용하여 회귀분석과 시계열 분석을 수행하였다.

회귀분석에는 XGBoost와 LightGBM 알고리즘을, 시계열 분석에는 LSTM 모델을 적용하여 예측 성능을 비교 분석하였다.

회귀분석 실험 결과, 전체 특성을 사용한 경우와 상관관계 분석을 통해 선별된 특성을 사용한 경우 모두 음수의 R^2 값을 기록하였다. 이는 모델의 예측력이 종속변수의 평균값을 사용하는 기준 모델보다 낮은 수준임을 의미한다. 구체적으로 XGBoost 모델은 RMSE 0.799, MAE 0.391의 성능을 보였으며, LightGBM 모델은 RMSE 0.582, MAE 0.346을 기록하였다. 이러한 회귀모델의 낮은 성능은 토마토 당도가 단일 시점의 환경 요인이나 생육 특성만으로는 설명하기 어려운 시계열적 특성을 내포하고 있음을 시사한다.

반면, LSTM을 활용한 시계열 분석에서는 MSE 0.086, RMSE 0.292, MAE 0.224, R^2 0.588의 통계적으로 유의미한 성능 향상을 달성하였다. MAE 0.224는 예측값과 실제 당도 간 평균적으로 0.22 Brix의 절대 차이가 존재함을 의미하며, 이는 토마토 당도의 일반적 범위인 3~5 Brix 대비 약 5.5%의 상대 오차율에 해당한다. LSTM 모델이 회귀분석 대비 우수한 성능을 보인 것은 시계열 데이터의 장기적 패턴과 시간적 의존성을 효과적으로 포착할 수 있는 모델 구조에 기인한다.

LSTM 모델의 R^2 0.588이 나타내는 중간 수준의 설명력은 다음과 같은 기술적 한계에서 기인하는 것으로 분석된다. 첫째, 5분 단위의 센서 데이터와 1주 단위의 생육 데이터 간 시간적 해상도의 불일치가 존재한다. 이는 데이터 수집 주기의 최적화와 해상도 통일을 통해 개선이 가능하다. 둘째, 주차별 평균화 과정에서 발생하는 정보 손실이 모델의 성능을 제한한다. 이는 시계열 데이터의 특성을 보존하는 고도화된 전처리 기법의 도입을 통해 최소화할 수 있다. 셋째, 토마토 생육 과정의 비선형적 시계열 특성을 모델이 완전히 포착하지 못하는 한계가 있다. 이는 당도 형성에 영향을 미치는 추가적 환경 요인 및 생육 특성 변수의 도입을 통해 보완이 가능하다. 이러한 체계적 개선을 통해 모델의 예측 성능과 설명력의 향상이 가능할 것으로 분석된다.

본 연구는 비파괴 방식의 실용적 당도 예측 가능성을 실증적으로 입증하였다는 점에서 학술적 성과와 실용적 활용성을 높인다는 점에서 의의가 있다. LSTM 모델이 달성한 58.8%의 설명력과 5.5%의 상대 오차율은 기존 파괴검사 방식을 대체할 수 있는 정량적 근거를 제시한다. 본 연구 결과는 토마토 당도 예측에 있어 시계열 분석 기법의 효용성을 입증했다.

감사의 글

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-2020-001489).

References

[1] A. Kamilaris and F. X. Prenafeta-Boldu, "Deep learning in agriculture: A survey," *Computers and electronics in agriculture*, vol. 147, 2018, pp. 70-90. https://doi.org/10.1016/j.compag.2018.02.016.

[2] S. Jung, "A Study on Consumer Purchasing Behavior of Imported and Produced Fruits in Korea", *Master's Thesis, School of Graduate Jeonbuck National University*, 2022

[3] S. A. Jaywant, H. Singh, and K. M. Arif, "Sensors and Instruments for Brix Measurement: A Review," *Sensors*, vol. 22, no.6, 2022, pp.2290. https://doi.org/10.3390/s22062290

[4] D. Lee, J. Eom, "Implemented of non-destructive intelligent fruit Brix(sugar content) automatic measurement system," *J. of sensor science and technology*, vol. 29, no. 6, pp.433-439, 2020. https://doi.org/10.46670/JSST.2020.29.6.433

[5] A. Hong, D. Noh, J. Choi, "A Study on the Correlation between Smart Farm Tomato Yield and Growth Data Using Multiple Regression Analysis," *In Proc. of Symposium of the Korean*

Institute of communications and Information Sciences, June 2022, pp.0763-0764.

[6] H. Oh, J. Lim, S. Yang, Y. Cho, and C. Shin, "A Study on the Prediction of Strawberry Production in Machine Learning Infrastructure," *J. of Smart Medial*, vol.11, no.5, 2022, pp.9-16. https://dx.doi.org/10.30693/SMJ.2022.11.5.9

[7] S. Kim, S. Kim, G. Hahn, I. Choi, J. Kyung, E. Lee, and Y. Choi, "Smart Farm Strawberry Production Forecast Model: Focusing on Variable Selection for Time Series Data," *In Proc. of KIIT Conference*, November 2023, pp.744-748

[8] B. Jo, S. Lee, and K. Kim, "Determination of SSC in Oriental Melon Using Snapshot-type Hyperspectral Imagery," *In Proc. of The Institute of Control, Robotics and Systems Conference*, Jeonnam, Korea, June 2022, pp.497-498

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, 1997, pp. 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] S. Lee, H. Yang, M. Kim, J. Kim, A. Son, and S. Hong, "A Study on Tomato Yield Prediction Based on BI-LSTM for Smart Farm Applications," *J. of the Korea Institute of Communication Sciences*, vol. 48, no. 4, 2023, pp. 457-468.

[11] H. Yu, H. Kim, and J. Moon, "Utilizing LSTM Networks for Multivariate Air Pollution Forecasting," *In Proc. of The Korean Society of Computer Information Conference*, July 2024, pp.877-878.

[12] H. Kim, H. Yu, and J. Moon, "Agricultural Meteorological Data Analysis for Real-Time Soil Moisture Measurement Using LSTM," *In Proc. of The Korean Society of Computer Information Conference*, July 2024, pp.693-64.

[13] AI Hub. Integrated Smart Farm Data (Tomato). Korea Artificial Intelligence Hub, 2022. Available at: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=534.

저자 소개

백남곤(Nam-Gon Beak)

2009년 국립순천대학교 정보통신공학과 졸업(공학사)
2023년 ~ 현재 국립순천대학교 대학원 스마트융합전공 석사과정

※ 관심분야 : 시계열 데이터 분석, 데이터 마이닝, 스마트팜

김진성(Jin-Sung Kim)

2021년 국립순천대학교 컴퓨터공학과 졸업(공학사)
2023년 국립순천대학교 멀티미디어공학과 졸업(공학석사)

2023년 ~현재 국립순천대학교 대학원 IT융합바이오융합시스템전공 박사과정
※ 관심분야 : 컴퓨터비전, 딥러닝, 시맨틱 세그멘테이션, 빅데이터 분석, 머신러닝

최은성(Eun-Sung Choi)

2024년 국립순천대학교 무역학과 졸업(경영학사, 공학사)
2024년~ 현재 국립순천대학교 컴퓨터공학과 석사과정 재학

※ 관심분야 : 컴퓨터비전, 딥러닝, 시계열 데이터, 설명 가능한 인공지능

심춘보(Chun-Bo Sim)

1996년 전북대학교 컴퓨터공학과 졸업(공학사)
1998년 전북대학교 대학원 컴퓨터공학과 졸업(공학석사)

2003년 전북대학교 대학원 컴퓨터공학과 졸업(공학박사)
2005년 ~현재 순천대학교 인공지능공학부 교수
※ 관심분야 : 빅데이터, 블록체인, 딥러닝, 생성모델, 자연어 처리, 강화학습

정세훈(Se-Hoon Jung)

2012년 순천대학교 대학원 멀티미디어공학과 졸업(공학석사)
2017년 순천대학교 대학원 멀티미디어공학과 졸업(공학박사)
2018년 영산대학교 빅데이터융합전공 조교수
2020년 안동대학교 창의융합학부 조교수
2022년 ~현재 순천대학교 컴퓨터공학과 부교수
※ 관심분야 : 블록체인, 딥러닝, 빅데이터 분석