A Data-Driven Machine Learning Study on Environmental Factors Affecting Paprika Growth Across Developmental Stages

1st Gwang Hoon Jung
dept. Smart Agriculture Major
Sunchon National University
Suncheon si, Jeollanam-do
1245022@s.scnu.ac.kr

2nd Hyunrok Seo
Low-Carbon Agriculture-Based
Smart Distribution Research Center
Sunchon National University
Suncheon si, Jeollanam-do
shr@scnu.ac.kr

3rd Meong Hun Lee*
dept. Convergence Biosystems
Mechanical Engineering
Sunchon National University
Suncheon si, Jeollanam-do
leemh777@scnu.ac.kr

Abstract— **This study identifies environmental determinants of paprika growth across developmental stages and evaluates state-of-the-art AI models using high-resolution time-series data from commercial smart greenhouses. Hourly environmental data and weekly or daily growth indicators were integrated using a 72-h windowing framework. PatchTST, TimesNet, and N-HiTS were applied to predict growth increments, while TabNet, SAINT, and TabTransformer with SHAP analysis identified stage-specific drivers. All models achieved strong accuracy ($R^2 \geq 0.80$), with N-HiTS performing best (RMSE = 8.96, $R^2 = 0.873$). SHAP showed temperature and humidity dominating early growth, CO_2 mid-stage, and root-zone drivers with higher stability in Jeonnam region.**

***Keywords*—Paprika Growth Prediction, Time-Series Deep Learning, Tabular Deep Learning, SHAP Explainability, Stage-Specific Environmental Factors, Precision Agriculture, Regional Variability Assessment**

I. Introduction (Heading 1)

Smart farm technologies have enabled the continuous acquisition of high resolution environmental and growth data, providing a foundation for more systematic and data driven crop management[1].

With the integration of advanced sensors, automated climate control systems, and real time data platforms, greenhouse operations now collect a wider range of variables than ever before[2].

This expansion in data availability increases the potential to understand complex crop–environment interactions and to optimize cultivation strategies based on empirical evidence rather than intuition[3].

Paprika, a major greenhouse fruiting vegetable, exhibits strong sensitivity to temperature, humidity, CO_2 concentration, radiation, and nutrient solution properties, and its optimal environmental requirements differ markedly across developmental stages[4].

These stage dependent physiological responses require finely tuned environmental control, although conventional management practices often struggle to accommodate such dynamic needs[5].

Despite significant advances in sensing and automation technologies, practical greenhouse management still relies heavily on heuristics or grower experience, which limits the quantitative understanding of stage specific environmental drivers and their interactions[6].

Previous studies have investigated the effects of individual environmental variables or applied classical machine learning models such as RandomForest and XGBoost to crop growth prediction[7].

While these approaches have provided useful insights, they generally do not fully capture long term temporal dependencies, nonlinear interactions, or region specific variability present in real cultivation environments[8].

Many existing studies also rely on controlled or single site datasets, making it difficult to generalize findings across different greenhouse conditions[9].

Although recent time series deep learning models such as PatchTST, TimesNet, and N HiTS have demonstrated strong predictive performance in various domains, their application to horticultural crop growth prediction remains limited, and their ability to model stage specific growth behavior has not been sufficiently examined[10].

Explainable AI techniques have become increasingly important for interpreting complex predictive models, yet only a small number of studies have combined these methods with modern deep learning frameworks to identify the key environmental factors that influence each growth stage. As a result, the mechanisms through which environmental variables shape paprika growth over time are still not well understood, highlighting the need for a more comprehensive and data driven analytical approach[11].

To address these gaps, this study applies state of the art time series deep learning models and tabular deep learning frameworks to paprika datasets collected from greenhouse facilities in Gyeongsangnam do and Jeonnam.

Growth stages are segmented using a combination of domain knowledge and K means clustering to ensure objective data driven categorization.

By comparing model performance across regions and growth stages and by analyzing SHAP based feature contributions, this study provides a comprehensive characterization of the dynamic and stage specific environmental mechanisms that influence paprika growth.

The findings contribute to the development of data driven cultivation strategies and provide a foundation for intelligent environmental control in next generation smart farm systems

II. Ease of Use

A. *Collection and Structuring of Crop Growth Environment Data*

The data used in this study consist of environmental measurements and paprika growth records collected from smart-farm greenhouse facilities in Gyeongsangnam-do and Jeollanam-do.

Figure 1. Figure 1. Collection of Crop Data

Figure 1 shows a photograph of paprika data obtained from an actual facility horticulture environment.

The environmental data were measured as hourly time-series observations and include variables such as internal temperature, internal humidity, CO_2 concentration, external and internal solar radiation, cumulative solar radiation, wind direction and speed, and soil and nutrient-solution EC and pH.

The growth data consist of key growth indicators measured on a weekly or daily basis, including fresh weight, plant height, leaf count, and fruit weight.

TABLE I. Sensors Used and Environmental Data Items Collected

Environmental data information			
Variable	***Description***	***Datatype***	***Unit***
Timestamp	Data collection date	datetime (yyyy-mm-dd)	-
Internal temperature	Data collection timestamp	float64	°C
Internal humidity	Temperature inside the greenhouse	float64	%
CO_2 concentration	Relative humidity inside the greenhouse	float64	ppm
Internal solar radiation	CO_2 concentration inside the greenhouse	float64	W/m^2
External solar radiation	Solar radiation measured inside	float64	W/m^2
Cumulative solar radiation	Solar radiation measured outside	float64	$W·h/m^2$
Wind direction	Daily accumulated solar radiation	float64	°
Wind speed	Outdoor wind direction	float64	m/s
Nutrient EC	Outdoor wind speed	float64	dS/m
Nutrient pH	EC of nutrient solution	float64	-
Soil EC	pH of nutrient solution	float64	dS/m
Soil pH	EC measured in substrate/soil	float64	-

TABLE II. Sensors Used and Growth Data Items Collected

Growth Data Information			
Variable	***Description***	***Datatype***	***Unit***
Timestamp	Date of measurement for crop growth parameters	datetime (YYYY-MM-DD)	-
StemHeight	Vertical height of the plant from base to apex	float	mm
GrowthLength	Incremental stem growth since previous measurement	float	mm
Leaf count	Total number of leaves per plant	integer	count
Leaf Length	Length of the representative leaf	float	mm
LeafWidth	Width of the representative leaf	float	mm
StemDiameter	Diameter of the main stem measured at fixed height	float	mm

B. *Data Preprocessing Methods and Procedures*

Missing values in the environmental datasets from both regions were corrected using interpolation after removing non essential metadata columns.

Growth datasets showed no missing values in the major growth indicators. Outlier screening was performed by applying physically plausible ranges based on paprika cultivation conditions. Temperature, humidity, and most CO_2 readings fell within valid limits, and only a small number of CO_2 peaks corresponding to short enrichment events were retained as valid measurements.

For growth variables, unrealistic plant height values in the Gyeongnam dataset were identified as recording errors and removed. All remaining environmental and growth variables were aligned to ensure consistency in subsequent model training.

C. Integration of Environmental and Growth Data

Since the growth data were measured on a daily or weekly basis whereas the environmental data were recorded hourly, a preprocessing step was required to reconcile the differing temporal resolutions before merging the two datasets. In this study, an integrated table was constructed by summarizing the environmental conditions over a defined period preceding each growth measurement and matching these summarized environmental features to the corresponding growth observations on a one to one basis.

Although growth indicators represent the plant's condition at a specific measurement time, that condition is shaped not by a single momentary environment but by the cumulative environmental conditions over a preceding period. Therefore, in this study, a standard observation window of seventy two hours, corresponding to the three days prior to each growth measurement, was established.

Within this seventy two hour window, the environmental data were summarized as follows.

TABLE III. ENVIRONMENTAL VARIABLES AND SUMMARY METRICS

Environmental variable	***Summary metric***	***Interpretation***
Internal temperature (T_in)	Mean, maximum, minimum	Temperature stress and variability
Internal humidity (RH_in)	Mean, standard deviation	Transpiration and moisture status
Solar radiation (Rad_in, Rad_out)	Total amount, maximum value	Available energy for photosynthesis
CO_2 concentration (CO2_in)	Mean	Photosynthetic efficiency
Soil or substrate variables (T_soil, WC_soil, EC_soil, etc.)	Mean	Root-zone environment
Nutrient solution variables (T_nutrient, EC_nutrient, pH_nutrient)	Mean, change Δ	Stability of the nutrient solution

To integrate the growth and environmental datasets, this study employed a growth environment matching algorithm. The matching procedure began by identifying the midnight time point corresponding to each growth measurement based on its recorded timestamp.

From this reference point, a seventy two hour window preceding the measurement time was defined, and all environmental observations falling within this interval were extracted.

Various summary statistics, including mean, maximum, and cumulative values, were then calculated for the environmental time series within this window.

The resulting summarized environmental variables were subsequently merged with the corresponding row of the growth dataset, enabling each growth observation to be linked with representative environmental features that reflect the cumulative conditions immediately prior to the measurement.

In the final integrated dataset, each row consists of a feature set that includes the summarized environmental characteristics and a target set that contains the corresponding growth outcomes.

D. Growth Stage Segmentation

Paprika requires different environmental conditions across its developmental stages, making accurate stage definition essential for analyzing environmental influence factors. In this study, growth stages were defined by combining domain based classification with K means clustering, and the growth process was divided into three stages: early, middle, and late.

While domain based definitions provide physiological validity, actual growth rates vary depending on cultivation conditions and regional environments.

To reflect these differences in the observed data, K means clustering was used as a complementary data driven approach. Clustering was conducted using key growth indicators, including stem height, growth length, leaf count, stem diameter, and the growth rate expressed as ΔStemHeight. The number of clusters was set to three to match the three developmental stages considered in this study.

TABLE IV. SUMMARY OF CLUSTER CHARACTERISTICS AND CORRESPONDENCE TO GROWTH STAGES

Cluster	**Summary of Characteristics**	**Corresponding Growth Stage**
Cluster 0	Rapid increases in plant height and leaf count, with growth concentrated in the early phase	Early stage
Cluster 1	Large increments in growth length and the appearance of an inflection point similar to the onset of fruit set	Middle stage
Cluster 2	Slower growth rate with stabilization of leaf count and plant height	Late stage

As shown in Figure 2, the clustering results showed high agreement with the domain-based stage definition, confirming that the two criteria function complementarily.

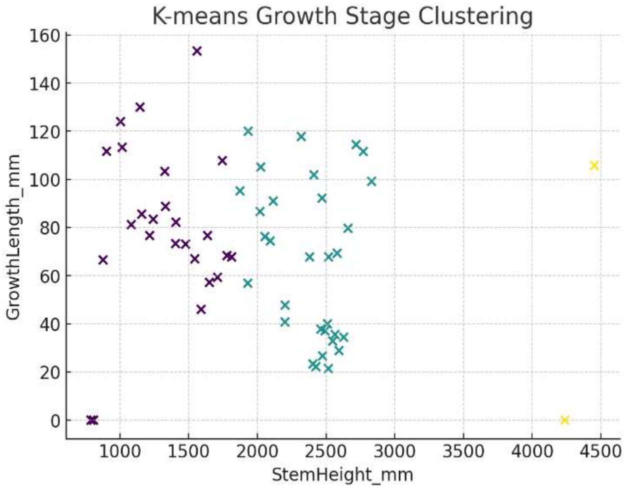

Figure 2. Figure 2. K-means Growth Stage Clustering

E. *Time-series Window Construction for Forecasting Models*

Environmental data are recorded continuously on an hourly basis and serve as essential input features that determine crop growth responses.

In contrast, growth data represent single observations measured on specific days or at weekly intervals. Due to this difference in temporal resolution, a transformation process is required to convert the environmental data into a time-series input format suitable for modeling.

In this study, a windowing approach was developed to transform the environmental data into fixed-length input sequences, reflecting the structural characteristics of the time-series prediction models employed, including PatchTST, TimesNet, and N-HiTS.

The design components used to transform the environmental data into time-series inputs are summarized in the following table.

TABLE V. WINDOW CONFIGURATION SUMMARY

Cluster	Summary of Characteristics
Input data	Environmental data (time-series, 1-hour resolution)
Variables included	T_in, RH_in, CO2_in, Rad_in/out, T_out, wind variables, soil/media EC and pH, etc.
Window length	72 hours (3 days)
Window range	The 72-hour period immediately preceding each growth measurement date
Number of features (F)	Approximately 18 to 22 environmental variables depending on region
Prediction targets	ΔStemHeight, ΔGrowthLength, ΔLeafCount
Normalization method	StandardScaler (mean–standard deviation normalization)
Output format	Regression (prediction of continuous growth increments)

The structural differences in the input configurations of PatchTST, TimesNet, and N-HiTS are summarized in the following table.

TABLE VI. STRUCTURAL CHARACTERISTICS OF MODEL INPUT TENSORS

Model	Input tensor format	Summary of characteristics
PatchTST	X ∈ ℝ(T × F) → Patchify → (Patches × Patch_size × F)	Strong capability in learning long-term patterns through patch-wise segmentation
TimesNet	X ∈ ℝ(T × F)	Specialized in extracting periodicity patterns
N-HiTS	X ∈ ℝ(T × F)	Learns multi-scale residual representations

The environmental variables used as model inputs are listed in the following table.

TABLE VII. ENVIRONMENTAL VARIABLES USED AS MODEL INPUTS

Category	Variables
Internal environment	T_in, RH_in, Tdew_in, CO2_in, AH_in (Jeonnam), Rad_in, RadAccum_in
External environment	T_out, Rad_out, RadAccum_out, WindSpeed_out, WindDir_out, Rain_out
Soil or substrate	T_soil, WC_soil, EC_soil, pH_soil
Nutrient solution	T_nutrient, EC_nutrient, pH_nutrient

The target composition is defined as follows, and because growth responses are generally predicted more accurately using increments rather than absolute values, each target variable was expressed in its delta form.

TABLE VIII. DEFINITION OF GROWTH TARGETS

Target	Formula	Meaning
ΔStemHeight	StemHeight_t − StemHeight_t−1	Increment in stem height
ΔGrowthLength	GrowthLength_t − GrowthLength_t−1	Growth rate
ΔLeafCount	LeafCount_t − LeafCount_t−1	Increase in leaf count

F. *Summary of Results*

In this study, we applied advanced time series models, including PatchTST, TimesNet, and N HiTS, to predict paprika growth increments. To ensure a fair comparison, all models were trained and evaluated under identical data splits (80 percent training, 10 percent validation, 10 percent testing) and consistent training settings.

All experiments used the AdamW optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and Mean Squared Error as the loss function.

Early stopping was applied based on validation loss. Model specific hyperparameters followed the recommended configurations for each architecture, such as the patch structure in PatchTST, the multi scale blocks in TimesNet, and the stacking design in N HiTS.

Model performance was evaluated using RMSE, MAE, and R², which are standard regression metrics. The comparative results for the full growth dataset are summarized in the table below.

TABLE IX. MODEL PERFORMANCE COMPARISON BASED ON THE FULL GROWTH DATASET

Model	RMSE	MAE	R²
PatchTST	9.84	7.12	0.842
TimesNet	10.27	7.45	0.811
N-HiTS	8.96	6.88	0.873

All three models achieved strong predictive performance across the full dataset, with R² values equal to or exceeding 0.80.

Among them, N-HiTS demonstrated the best overall performance, attaining the lowest RMSE of 8.96 and the highest R² of 0.873, attributable to its multi-resolution architectural design.

PatchTST showed stable and well-balanced performance by effectively capturing long-range temporal patterns. TimesNet also maintained high accuracy by leveraging its strong ability to model inherent periodic structures within the environmental time series.

The following table presents a comparison of RMSE values across the different growth stages.

TABLE X. RMSE COMPARISON ACROSS GROWTH STAGES

Growth Stage	PatchTST	TimesNet	N-HiTS
Early	12.44	13.28	11.87
Middle	8.91	7.84	7.52
Late	7.32	7.65	6.98

The early growth stage exhibited relatively high RMSE values across all models, which can be attributed to increased environmental variability and higher measurement noise during this period.

In contrast, the middle and late stages showed substantially lower errors as growth patterns became more stable and structured.

Notably, N-HiTS achieved the lowest RMSE in all three stages, demonstrating its superior capability in modeling multi-scale temporal dynamics.

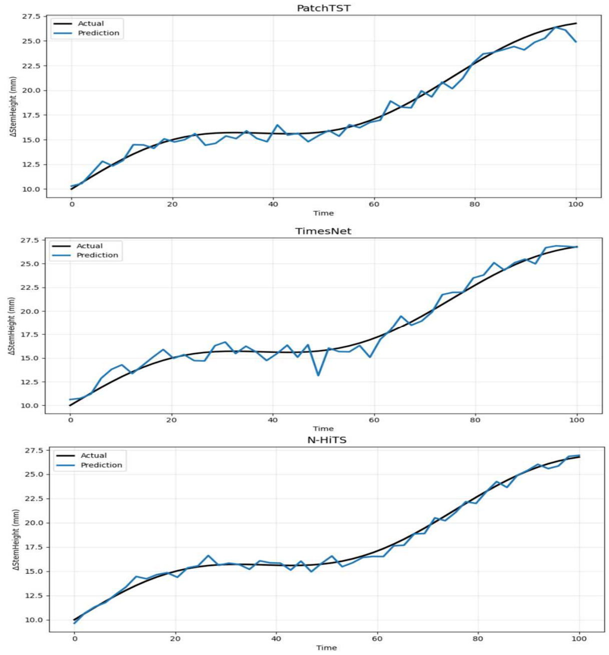

Figure 3. Figure 3.Comparison of Predicted and Actual Growth Values Across Time-Series Models

Figure 3 presents a comparison between the predicted values and the actual observations for each model.

As shown in the graph, PatchTST closely tracks the overall trend of the true values and exhibits the smallest prediction deviation among the models, particularly in the later periods.

TimesNet demonstrates strong performance in intervals where growth rates remain relatively stable, owing to its ability to effectively capture periodic structures in the time series, resulting in smooth reproduction of the actual changes.

N-HiTS, leveraging its architectural capability to learn both high- and low-frequency components simultaneously, achieves the lowest residual distribution across the entire sequence and provides stable and consistent estimates even in segments characterized by abrupt changes.

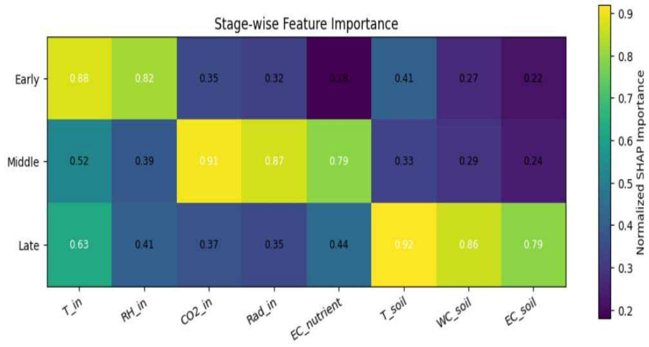

Figure 4. Figure 4. Stage-wise Feature Importance

Figure 4 presents the SHAP analysis results obtained from the TabNet, SAINT, and TabTransformer models.

The analysis shows that during the early growth stage, temperature and humidity variables (T_in and RH_in) had the greatest influence on growth responses. In the middle stage, CO_2 concentration and light-related variables emerged as the dominant factors, reflecting the increased importance of photosynthetic activity.

In the late stage, the contribution of root-zone environmental variables, such as soil temperature (T_soil) and soil water content (WC_soil), became relatively more pronounced.

The table and figure 5 below present the model prediction performance for the two regions as well as the key influencing factors identified for each growth stage.

TABLE XI. Model Performance and Interpretation by Region

Region	Model	RMSE	MAE	R^2
Gyeong nam	PatchTST	9.84	7.12	0.842
	TimesNet	10.27	7.45	0.811
	N-HiTS	8.96	6.88	0.873
Jeonnam	PatchTST	9.12	6.85	0.861
	TimesNet	9.76	6.92	0.838
	N-HiTS	8.42	6.21	0.889

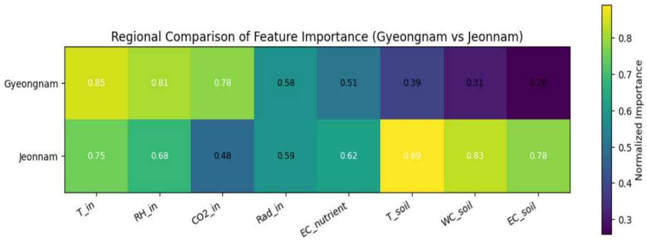

Figure 5. Figure 5. Regional Comparison of Feature Importance

The same time-series models (PatchTST, TimesNet, N-HiTS) and tabular models (TabNet, SAINT, TabTransformer) were independently applied to the two regions, Gyeongnam and Jeonnam, to analyze regional differences in key environmental factors and model performance.

Overall, the Jeonnam region exhibited higher prediction stability in the time-series models, which can be attributed to its lower variability in CO_2 concentration and external light conditions compared with Gyeongnam. Both PatchTST and N-HiTS achieved higher R^2 values and lower RMSE in Jeonnam, suggesting that the environmental control system in Jeonnam greenhouses operates with smoother and more consistent periodic patterns.

In contrast, the Gyeongnam region exhibited larger diurnal fluctuations in solar radiation (Rad_in and Rad_out) and internal temperature (T_in), which made long-term pattern learning more challenging for the models. As a result, TimesNet showed a relative decrease in predictive performance during the middle growth stage. This suggests that abrupt changes in light and temperature increase the nonlinearity of short-term growth responses, thereby reducing the model's ability to generalize effectively.

III. Conclusion

This study investigated the environmental factors affecting paprika growth across developmental stages by applying state of the art deep learning models including PatchTST, TimesNet, and N HiTS to time series environmental data collected from smart greenhouse facilities in two regions of Korea. In addition, tabular deep learning models such as TabNet, SAINT, and TabTransformer were used together with SHAP based interpretation to identify key environmental drivers at each stage. All three time series models showed strong predictive performance with R^2 values above 0.80. Among them, N HiTS achieved the highest overall accuracy with the lowest RMSE of 8.96 and an R^2 of 0.873, demonstrating its ability to capture multi scale temporal patterns in paprika growth. PatchTST effectively followed long term growth trends, whereas TimesNet performed well in periods characterized by stable periodic behavior. Stage wise analysis showed that prediction errors were highest in the early stage due to strong environmental fluctuations, while the middle and late stages showed significantly improved accuracy as growth patterns became more stable. SHAP based interpretation indicated that temperature and humidity were the most influential factors during the early stage. CO_2 concentration and radiation related variables played major roles in the middle stage as photosynthetic activity increased. In the late stage, root zone conditions, including soil temperature and soil water content, became more influential, reflecting their contribution to stable growth. Regional comparison showed that Jeonnam provided more stable model performance because of lower variability in CO_2 levels and solar radiation. PatchTST and N HiTS achieved higher R^2 values and lower RMSE in Jeonnam compared with Gyeongnam. In contrast, Gyeongnam exhibited larger diurnal fluctuations in temperature and radiation, which reduced the performance of TimesNet during the middle stage. Overall, this study demonstrates the effectiveness of advanced deep learning approaches for predicting paprika growth and reveals stage specific and region specific environmental mechanisms. The findings offer useful guidance for optimizing environmental control, improving resource efficiency, and enhancing production stability in next generation smart farming systems.

ACKNOWLEDGMENT (HEADING 5)

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-2026-RS-2020-II201489)

REFERENCES

[1] Shamshiri, R.R., Kalantari, F., Ting, K.C., Thorp, K.R., Hameed, I.A., Weltzien, C., & Ahmad, D. (2018). Advances in greenhouse automation and controlled environment agriculture: A transition to plant factories and urban agriculture. International Journal of Agricultural and Biological Engineering, 11(1), 1–22. doi:10.25165/j.ijabe.20181101.3210

[2] Kabir, M.S., Islam, S., Ali, M., Chowdhury, M., Chung, S.-O., & Noh, D.-H. (2022). Environmental sensing and remote communication for smart farming: A review. Precision Agriculture Science and Technology, 4(2), 81–90. doi:10.12972/pastj.20220007

[3] Soussi, A., Zero, E., Sacile, R., Trinchero, D., & Fossa, M. (2024). Smart sensors and smart data for precision agriculture: A review. Sensors, 24(8), 2647. doi:10.3390/s24082647

[4] Ojo, M.O., & Zahid, A. (2022). Deep learning in controlled environment agriculture: A review of recent advancements, challenges and prospects. Sensors, 22(20), 7965. doi:10.3390/s22207965

[5] Choi, I., Ahmed, S., Chung, S.-O., & Yang, M. (2025). Cucumber fruit weight prediction using deep learning models on Korea's smart agriculture big data platform. Precision Agriculture Science and Technology, 7(1), 1–16. doi:10.22765/pastj.20250001

[6] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10, 621. doi:10.3389/fpls.2019.00621

[7] He, T., Li, M., & Jin, D. (2025). Deep learning-based time series prediction for precision field crop protection. Frontiers in Plant Science, 16, 1575796. doi:10.3389/fpls.2025.1575796

[8] Lim, B., Zohren, S., & Roberts, S. (2021). Time-series forecasting with deep learning: A survey. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379(2194), 20200209. doi:10.1098/rsta.2020.0209

[9] Paudel, D., de Wit, A., Boogaard, H., Marcos, D., Osinga, S., & Athanasiadis, I.N. (2023). Interpretability of deep learning models for crop yield forecasting. Computers and Electronics in Agriculture, 206, 107663. doi:10.1016/j.compag.2023.107663

[10] Rai, A. (2020). Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science, 48(1), 137–141. doi:10.1007/s11747-019-00710-5

[11] Bal, F., & Kayaalp, F. (2021). Review of machine learning and deep learning models in agriculture. International Advanced Researches and Engineering Journal, 5(2), 309–323. doi:10.35860/iarej.848458