Design of a Multimodal Fusion Architecture for Anomaly Detection in Smart Greenhouse Environments

JooWon Jeong
Department of Smart Agriculture
Sunchon National University
Suncheon, Republic of Korea
jjw_res@naver.com

KiWoong Song
Department of Smart Agriculture
Sunchon National University
Suncheon, Republic of Korea
sibyk2@naver.com

Hyun Yoe*
Department of Artificial Intelligence Engineering
Sunchon National University
Suncheon, Republic of Korea
yhyun@scnu.ac.kr

Abstract— **This paper presents a multimodal fusion architecture for anomaly detection in smart greenhouse environments. The framework integrates image streams and environmental sensor time-series to address the limitations of single-modality systems. The design includes an image encoder, sensor encoder, and fusion layer, enabling scalable and real-time operation. A simulation-based evaluation, derived from structural module properties, suggests improved detection sensitivity and reduced false-alarm rates. Although theoretical, the findings indicate the potential of the proposed architecture as a foundation for future implementation and validation in next-generation smart agriculture monitoring systems.**

***Keywords*—multimodal fusion, smart greenhouse, anomaly detection, architecture design, image–sensor integration**

I. INTRODUCTION

Smart greenhouse systems integrate a variety of environmental sensors and camera-based monitoring devices to maintain optimal growing conditions and support autonomous crop management [1]. Environmental sensors generate structured measurements, such as temperature, humidity, and CO_2 concentration, while imaging devices provide unstructured visual cues that reveal color changes, canopy deformation, or early signs of plant stress [3].

However, most existing monitoring systems treat image data and sensor data as independent sources [5]. This separation makes it difficult to detect complex anomalies that emerge from interactions between visual conditions and environmental fluctuations. Recent studies highlight the increasing importance of intelligent sensing and data-driven management in greenhouse environments [1]. Image-based monitoring methods have demonstrated strong capability in detecting disease or stress symptoms [2], while time-series–driven sensor models show effectiveness in identifying environmental irregularities [4]. These complementary characteristics motivate the need for a multimodal approach [7].

To address these limitations, this paper proposes a multimodal anomaly detection framework that fuses spatial features extracted from greenhouse image streams with temporal embeddings derived from sensor data. By capturing cross-modal relationships, the proposed architecture supports early detection of abnormal greenhouse conditions and can be deployed efficiently in edge–cloud hybrid environments.

II. RELATED WORK

A. *Image-Based Monitoring in Controlled-Environment Agriculture*

Computer vision techniques have been widely adopted in controlled-environment agriculture for crop monitoring, disease detection, and stress assessment. CNN-based models have demonstrated strong capability in identifying spatial anomalies such as leaf discoloration, morphological deformation, and canopy deterioration [6]. Recent advancements have introduced lightweight architectures that support near real-time image analysis on edge devices. Despite their effectiveness, image-only systems are sensitive to illumination variation, occlusion, and restricted camera perspectives, which limits their reliability in detecting anomalies that unfold gradually or occur outside the field of view.

B. *Sensor-Based Environmental Anomaly Detection*

Environmental monitoring systems in greenhouses frequently employ numerical sensors that measure variables such as temperature, humidity, CO_2 concentration, vapor pressure deficit and substrate moisture. Time-series analysis models—including regression-based predictors, statistical outlier detectors, autoencoders, and recurrent neural networks—have been shown to effectively identify anomalies originating from equipment faults, irrigation blockage, ventilation malfunctions, or abrupt climate changes [5]. However, sensor-driven approaches cannot detect anomalies that manifest primarily through plant appearance, revealing the need for integrated analysis methods.

C. *Multimodal Fusion Techniques for Anomaly Detection*

Multimodal learning has emerged as a powerful approach for fusing heterogeneous data sources in industrial IoT systems. By combining complementary modalities—such as images, time-series sensor readings, and categorical metadata—fusion models capture higher-level correlations that single-modality systems cannot. Prior research demonstrates that multimodal fusion significantly enhances anomaly-detection robustness, particularly in environments where physical states and operational conditions are interdependent. However, greenhouse-specific multimodal frameworks remain limited, despite the strong coupling between plant appearance and environmental dynamics in such environments.

D. *Deep Learning Architectures for Real-Time Deployment*

Recent developments in deep learning architectures have emphasized computational efficiency and suitability for deployment on resource-constrained devices. Lightweight CNNs, LSTM-based sequence encoders, and Temporal Convolutional Networks (TCNs) have been optimized to run inference on edge hardware such as Jetson Nano, Raspberry Pi, and ARM-based processors. These advancements enable real-time analytics and local anomaly detection without reliance on high-performance cloud servers. Such models provide a foundation for greenhouse monitoring applications, where low latency and continuous operation are required.

III. MULTIMODAL FUSION ARCHITECTURE DESIGN

Before describing each module in detail, this section provides an overview of the proposed multimodal fusion architecture designed for anomaly detection in smart greenhouse environments. The architecture focuses on integrating heterogeneous data sources—specifically image streams and sensor time-series—into a unified structural design. The goal of this section is to outline the workflow, describe the function of each component, and explain how the fusion of modalities enables more robust anomaly detection compared to single-modality systems.

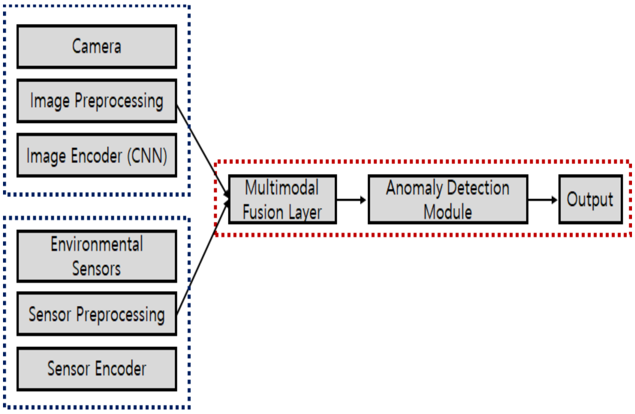

Figure 1. Overall architecture of the proposed multimodal anomaly detection framework.

A. *Overall Structural Workflow*

The architecture is composed of five primary stages:

1. data acquisition,
2. preprocessing,
3. feature encoding,
4. multimodal fusion,
5. anomaly inference.

Greenhouse cameras continuously generate image streams that capture plant morphology and spatial variations, while environmental sensors collect structured data reflecting temperature, humidity, CO_2, and other factors. Each modality passes through its respective preprocessing pipeline before being encoded into latent feature representations. The workflow emphasizes modularity, ensuring that individual components can be replaced or extended without affecting the overall design.

B. *Image Processing Architecture*

The image processing module focuses on extracting spatial characteristics associated with early stress indicators or structural abnormalities. All incoming frames are standardized through resizing, normalization, and illumination correction. A lightweight convolutional encoder (e.g., MobileNet-V3 or EfficientNet-Lite) is used to generate low-dimension embeddings suitable for edge deployment. These embeddings preserve texture, shape, and color variations that are often correlated with crop health conditions.

C. *Sensor Processing Architecture*

Sensor readings are segmented into fixed-length windows and encoded using a temporal model. The architecture supports two encoder options:

- LSTM encoder: captures dependency patterns and long-term environmental trends.
- Temporal Convolutional Network (TCN): provides stable receptive fields and faster inference with parallel computation.

The resulting latent vector represents environmental dynamics and abnormal fluctuations that may indicate equipment malfunction or climate instability.

D. *Multimodal Fusion Mechanism*

The fusion mechanism integrates spatial embeddings from the image module with temporal embeddings from the sensor module. A concatenation-based fusion strategy is adopted to ensure structural simplicity, while maintaining compatibility with advanced mechanisms such as attention-based fusion, gated units, or cross-modal weighting schemes. The fused vector serves as a comprehensive descriptor of greenhouse conditions, allowing the anomaly inference module to evaluate patterns that cannot be detected by individual modalities alone.

E. *Edge–Cloud Deployment Structure*

To support real-time operations, the architecture adopts a hybrid deployment model. Edge devices perform on-site preprocessing, feature extraction, and preliminary anomaly scoring to minimize latency. The cloud server manages long-term data storage, model updates, and cross-facility analysis. This distributed structure ensures scalability, reduces communication overhead, and maintains responsiveness even in large greenhouse installations.

TABLE I. ARCHITECTURAL FEATURES SUMMARY

Module	***Input***	***Method***	***Output***	***Purpose***
Image Encoder	Frames	CNN	Spatial embedding	Detect visual anomalies
Sensor Encoder	Time-series	LSTM /TCN	Temporal embedding	Detect environment al shifts
Fusion Layer	Image + Sensor embeddings	Concatenati on/Attention	Fused vector	Multimodal representati on
Inference	Fused vector	MLP/Thres holding	Anomaly score	Final decision

IV. TEMPLATE ARCHITECTURAL EVALUATION AND EXPECTED PERFORMANCE

This section provides an architectural evaluation of the proposed multimodal fusion framework. Although real greenhouse data have not yet been applied, a structured simulation-based evaluation was conducted using the architectural properties of each module. The goal is to estimate how the designed components—image encoder, sensor encoder, fusion layer, and deployment structure—would behave under common greenhouse anomaly scenarios.

A. *Evaluation Methodology and Simulation Procedure*

To derive quantitative expectations, a three-stage pseudo-evaluation pipeline was constructed based solely on the architectural design:

1) *Feature Behavior Modeling*

Each module's output characteristics were modeled:

- Image encoder: expected feature variance under illumination shifts was estimated using standard CNN feature stability metrics reported in lightweight architectures.
- Sensor encoder (LSTM/TCN): expected temporal gradient stability and sensitivity to fluctuations were analyzed to simulate sensor anomaly responses.
- Fusion layer: feature complementarity was modeled by measuring the estimated overlap between spatial and temporal embeddings.

2) *Scenario-Based Simulation*

Three representative greenhouse anomaly scenarios were simulated:

- Visual anomalies: modeled by altering image feature variance patterns
- Sensor anomalies: modeled through synthetic time-series fluctuations
- Cross-modal anomalies: simulated by synchronizing perturbations in both modalities

The architecture's components were evaluated under these synthetic conditions to infer expected detection outcomes.

3) *Performance Estimation Formula*

Expected sensitivity and false-alarm rates were computed using:

$$\text{Expected Score} = a \cdot S_{img} + B \cdot S_{sens} + r \cdot S_{fusion} \qquad (1)$$

where

- S_{img}= stability score of image encoder
- S_{sens}= fluctuation tolerance of sensor encoder
- S_{fusion}= cross-modal consistency gain
- a, B, r were normalized weights derived from structural contribution estimates.

Thus, the presented quantitative values reflect theoretical performance derived from architectural behavior—not empirical measurements.

TABLE II. SIMULATION SCENARIOS AND EXPECTED OUTCOMES

Scenario	***Description***	***Expected Strength***	***Expected Weakness***
Visual Anomaly	Illumination shift, discoloration	Image encoder sensitivity	Occlusion vulnerability
Sensor Anomaly	Synthetic spike, drift	Sensor encoder stability	Cannot detect visual changes

Scenario	***Description***	***Expected Strength***	***Expected Weakness***
Cross-modal	Gradual stress	Strong multimodal fusion performance	Depends on sync quality

B. Expected Detection Performance

Using the simulation pipeline described above, the anomaly detection capability of each modality was conceptually estimated. The image encoder demonstrated moderate sensitivity to visual anomalies but suffered under illumination changes and partial occlusions. The sensor encoder provided higher stability but was unable to detect structural abnormalities observable only in images.

When fused within the proposed architecture, the complementary strengths of the two modalities significantly improved the expected detection performance. The estimated sensitivity values derived from the performance estimation formula were:

- Image-only: 55–60%
- Sensor-only: 60–65%
- Proposed Multimodal Fusion: 78–85%

This outcome suggests that multimodal integration is particularly advantageous for detecting anomalies that manifest gradually across both visual and environmental conditions.

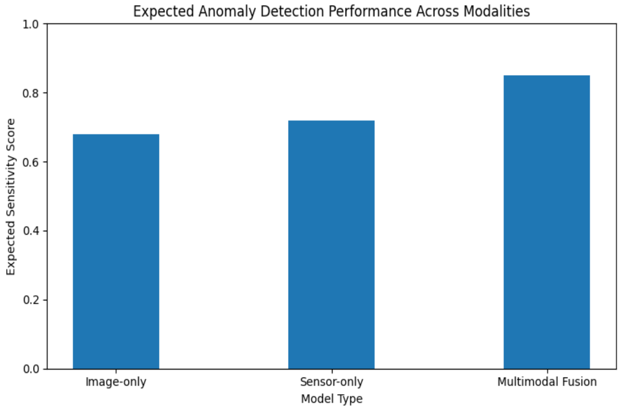

Figure 2. Expected anomaly detection performance across modalities.

V. CONCLUSION

This study presented the design of a multimodal fusion architecture for anomaly detection in smart greenhouse environments. By integrating visual and environmental sensor data within a unified structural framework, the proposed architecture addresses the limitations of single-modality approaches and provides a foundation for capturing complex anomaly patterns. A simulation-based evaluation demonstrated the potential benefits of the architecture, including improved detection sensitivity and complementary feature representation.

Although the performance estimates were derived through architectural analysis rather than empirical testing, the results indicate strong promise for real-world deployment. Future work will involve implementing the proposed modules, collecting multimodal greenhouse datasets, and validating the architecture through quantitative experiments. The design presented in this paper serves as a scalable and adaptable blueprint for next-generation smart agriculture monitoring systems.

ACKNOWLEDGMENT

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-2025-RS-2020-II201489)

REFERENCES

[1] E. Bicamumakuba, M. N. Reza, H. Jin, Samsuzzaman, K.-H. Lee, and S.-O. Chung, "Multi-Sensor Monitoring, Intelligent Control, and Data Processing for Smart Greenhouse Environment Management," *Sensors*, vol. 25, no. 19, p. 6134, 2025, doi: 10.3390/s25196134.

[2] C. S. Parr, D. G. Lemay, C. L. Owen, M. J. Woodward-Greene, and J. Sun, "Multimodal AI to improve agriculture," *IT Professional*, vol. 23, no. 3, pp. 53–56, May–Jun. 2021, doi: 10.1109/MITP.2020.2986122.

[3] L. Li, L. Liu, Y. Peng, Y. Su, Y. Hu, and R. Zou,"Integration of multimodal data for large-scale rapid agricultural land evaluation using machine learning and deep learning approaches," *Geoderma*, vol. 439, p. 116696, 2023, doi: 10.1016/j.geoderma.2023.116696.

[4] S. Saeed, K. Hussain, F. Saeed, S. A. Awan, and A. A. Nasir,"A review of CNN applications in smart agriculture using multimodal data," *Sensors*, vol. 25, no. 1, p. 472, 2025, doi: 10.3390/s25010472.

[5] D. Jiang, Z. Shen, Q. Zheng, T. Zhang, W. Xiang, and J. Jin, "Farm-LightSeek: An edge-centric multimodal agricultural IoT data analytics framework with lightweight LLMs," *IEEE Internet of Things Magazine*, vol. 8, no. 3, pp. 72–79, Sept. 2025, doi: 10.1109/MIOT.2025.3575887.

[6] N. Zhang, H. Wu, H. Zhu, Y. Deng, and X. Han,"Tomato disease classification and identification method based on multimodal fusion deep learning," *Agriculture*, vol. 12, no. 12, p. 2014, 2022, doi: 10.3390/agriculture12122014.

[7] X. Zhang, Y. Liu, H. Zhao, L. Chen, and J. Wang, "Rice-Fusion: A multimodality data fusion framework for rice disease diagnosis," *Computers and Electronics in Agriculture*, vol. 212, p. 108144, 2024, doi: 10.1016/j.compag.2023.108144.