Multimodal and Explainable AI System Design for Agricultural Sustainability: A Comprehensive Literature Review and Synthesis

Nilkanth Mukund Deshpande[1], Prof. Shashi Kant Gupta[1]

1. Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia

drnilkanthdeshpande@gmail.com, shashigupta@lincoln.edu.my

Abstract

Sustainability in agriculture is the key problem due to environmental issues, soil degradation and different crop diseases. Artificial Intelligence (AI) offer the solutions for precision agriculture with the utilization of multimodal systems. However, AI systems, especially utilizing deep learning are black boxes which will question the trust of diagnosis and predictions of the implemented system. This work explores a literature review of multimodal and explainable AI (XAI) systems for sustainable agriculture. It analyses multimodal fusion methodologies, deep learning, and explainable solutions that could be employed to assess sustainability impacts in agriculture. The study analyses the research prospects such as adaptive fusion, edge optimization, explainability evaluation.

Keywords: Multimodal AI, Explainable AI, Sustainable Agriculture, Precision Farming, Data Fusion, Crop Yield Prediction.

---

1. Introduction

Food is a necessity of the globe and there has always been the concern in regards to food security and sustainability [1]. This issue became critical due to the huge population of the globe[2]. In current agriculture the different multimodal challenges exist such as climate changes, degradation of soil, water stress, crop diseases, resource optimization[3].

These challenges are multimodal, since their analysis consists of different imaging techniques, sensor data, time series data of climate and inputs from farmers. Different AI systems could demonstrate the solid performance in the areas of crop disease detection, yield prediction, water-stress assessment, resource optimization

However, interpretability and explainability issues in the traditional deep learning models make the trust issues of stakeholders[4].

1.1 Motivation

Sustainable agricultural AI has to overlook the requirements including multimodal integration, predictive accuracy, explainability, and trustworthiness.

This paper analyses existing literature addressing these requirements.

2. Multimodal AI in Agriculture

2.1 Data Modalities

Recent literature integrates the following modalities with the application related to the agriculture

Table 1: Modalities of agricultural sustainability

Modality	**Examples**	**Application**
RGB imagery	UAV leaf images	Disease detection
Multispectral / Hyperspectral	Satellite bands	Crop stress
Thermal imaging	Canopy temperature	Water stress
Soil sensors	Moisture, pH	Irrigation planning

Modality	Examples	Application
Weather time series	Temperature, rainfall	Yield forecasting
Textual inputs	Farmer notes, Q/A	Advisory systems

2.2 Fusion Strategies

(A) Early Fusion (Feature-level)

In this method, the concatenation of features is done before the classification:
Following equation 1 shows the mathematics in this method.

$$Z = [X_{image} \oplus X_{sensor} \oplus X_{text}] \qquad (1)$$

where, Ximage, Xsensor, and Xtext are three different modalities .

(B) Late Fusion (Decision-level)

$$Y = \sum_{i=1}^{n} w_i \, f_i(X_i) \qquad (2)$$

where, Y represents the final prediction after fusion, n is the number of modalities, i represent the index of modality, $X_i$is the input from a particular modality, $f_i(X_i)$ is the prediction of model considering the modality of i, and $w_i$is the weight of the modality i.

(C) Hybrid / Adaptive Fusion

Attention-weighted fusion:

$$Z = \sum_{i=1}^{n} \alpha_i \, X_i \qquad (3)$$

where,
Z= final feature representation
α_i=attention weight
X_i=feature representation of individual modality i
N= number of different modalities.

$$\alpha_i = \frac{e^{e_i}}{\sum_j e^{e_j}} \qquad (4)$$

α_i=Softmax-normalized weight
e_iAttention score for modality i
Adaptive weighting improves cross-region generalization.

---

1. **Explainable AI in Agriculture**

Following figure 1 shows the different explainability in agriculture. As shown in the figure there are three broad types of explainability in agriculture. These includes post-hoc explainability, intrinsic explainability and human-centered explainability. Post-hoc type include the methods such as SHAP (shaply values), Local Interpretable Model-Agnostic Explanations (LIME), and Grad-CAM. Intrinsic methods includes attention-based transformers, prototype networks and decision trees for the explainability and interpretability. Moreover, the human-centered explanations also plays a major role here, including farmer comprehensions, their trust levels and actionability of farmers.

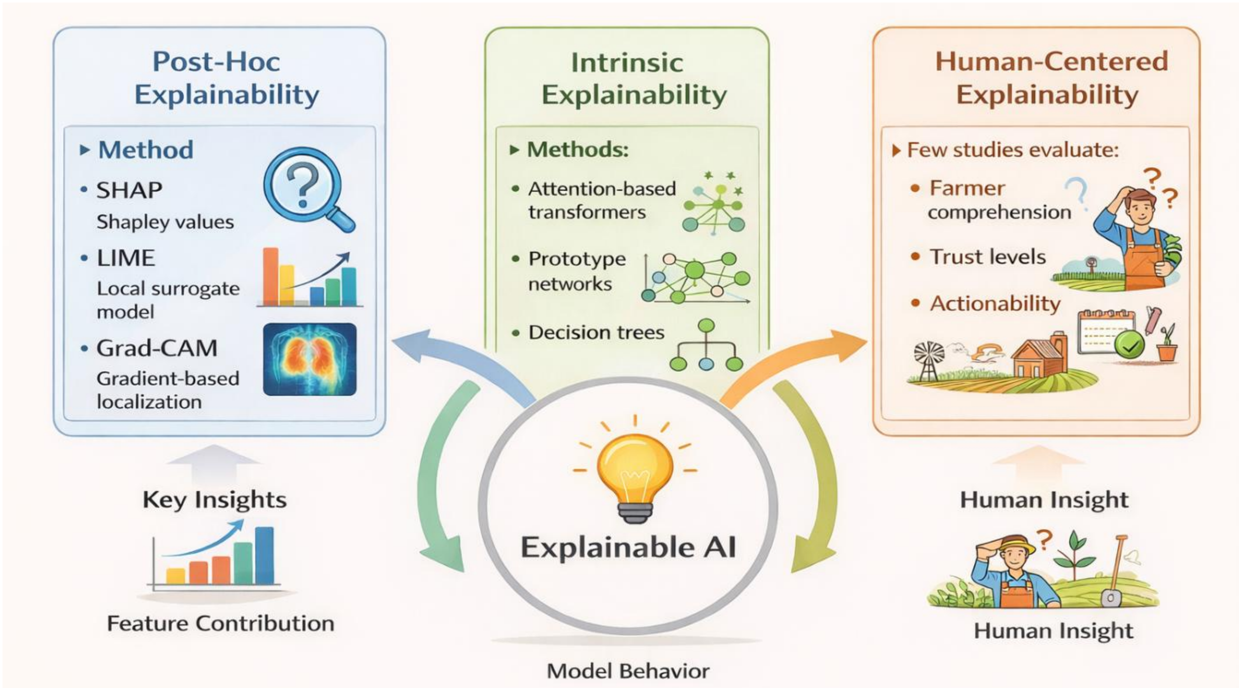

Figure: Different explainability types in agriculture

Table 2: Comparative Literature Summary

excel	Data Modalities	Method/ Model	Explainability Technique used	Research Findings
Zhou et al.[5]	Image + Metadata	CNN + rules	Grad-CAM	Accuracy improved
Feng et al.[6]	Image + Text	Cross-modal fusion	Attention	Reduced false positives
CropNet [7]	Satellite + Weather	ConvLSTM	Attention	Lower RMSE
Jiang et al.[8]	UAV + Weather	CNN Fusion	Grad-CAM	Improved yield prediction

5. Strengths of the existing research:

1. Stronger predictive performance
2. Emergence of multimodal benchmarks
3. Increasing XAI adoption
4. Cross-scale integration

6. Limitations and research gaps

Models fail across agro-ecological zones.
Attributions unstable and technical.
High computational cost.
Sensor failure affects predictions.

7. Research Directions

- Cross-region multimodal datasets
- Adaptive fusion under missing data
- Farmer-centered XAI evaluation
- Edge optimization & compression
- Federated learning for privacy

8. Conclusion

There are different methodologies for the agricultural sustainability including AI techniques utilizing machine learning, deep learning, and vision models. These techniques could provide good accuracies but lacks the interpretability and explainability concerns. To enhance the performance of AI systems in agriculture multimodal systems integration is required as most of the practical problems are multimodal in nature. Hence integration of different data streams will improve the performance of the system.
Future systems must integrate adaptive multimodal fusion with intrinsically interpretable architectures and rigorous socio-technical validation to ensure adoption in real-world agricultural ecosystems.

References

1. Berry, E. M., Dernini, S., Burlingame, B., Meybeck, A., & Conforti, P. (2015). Food security and sustainability: can one exist without the other?. *Public health nutrition*, *18*(13), 2293-2302.
2. Pinstrup-Andersen, P., Pandya-Lorch, R., & Rosegrant, M. W. (1999). World food prospects: Critical issues for the early twenty-first century.
3. Parr, C. S., Lemay, D. G., Owen, C. L., Woodward-Greene, M. J., & Sun, J. (2021). Multimodal AI to improve agriculture. *IT professional*, *23*(3), 53-57.
4. Triana-Martinez, J., Álvarez-Meza, A., & Castellanos-Dominguez, G. (2025). Enhancing agricultural data interpretability and visualization with TabNet-driven feature extraction and Local Biplots. *Results in Engineering*, 106672.
5. Zhang, Y., Zhu, Y., Liu, J., Yu, W., & Jiang, C. (2024). An interpretability optimization method for deep learning networks based on Grad-CAM. *IEEE Internet of Things Journal*, *12*(4), 3961-3970.
6. Feng, X., Zhao, C., Wang, C., Wu, H., Miao, Y., & Zhang, J. (2022). A vegetable leaf disease identification model based on image-text cross-modal feature fusion. *Frontiers in plant science*, *13*, 918940.
7. Lin, F., Guillot, K., Crawford, S., Zhang, Y., Yuan, X., & Tzeng, N. F. (2024). An open and large-scale dataset for multi-modal climate change-aware crop yield predictions.
8. Zhang, H., Zhao, S., Song, Y., Ge, S., Liu, D., Yang, X., & Wu, K. (2022). A deep learning and Grad-Cam-based approach for accurate identification of the fall armyworm (Spodoptera frugiperda) in maize fields. *Computers and Electronics in Agriculture*, *202*, 107440.