Design of a Multimodal Explainable AI Framework for Sustainable Intelligent Agriculture: Methodology and Expected Outcomes

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

Agricultural sustainability is a major concern currently in globe, as it is related to the food security and scarcity. There is always a need for multimodal and explainable frameworks for achieving the sustainability in the field of agriculture. The paper proposes a framework consisting of the integration of different modalities with explainable insights of the detection and prediction decision of crop management and precision agriculture.

A popular multimodal dataset Pheno4D is utilized for this work. The different modalities including RGB images, thermal images, 3D point cloud and time series data. These modalities offer a fair generalization of the proposed framework. The framework consists of a hybrid fusion architecture and attention-based model. Explainability and interpretability are achieved via attention mechanisms and post-hoc networks utilizing the frameworks including Shapley values (SHAP) and Grad-CAM. The proposed model primarily offers the disease prediction and its severity. The methodology could be extended for prediction of crop yield, and optimization of water-resources with maintaining the trust and interpretability.

1. Introduction

Agriculture is the backbone of the whole world, as it provides food, which is the most essential thing for the life of mankind. The sustainability of agriculture always demands accuracy, robustness, and transparency

This paper proposes a full-stack MM-XAI framework.

2. Brief Literature Review

As per the existing literature, the system should be multimodal [1] and accurate with the XAI enabling trust in the diagnosis and predictions [2]. The adaptive fusion could improve the robustness [3].

However, challenges of the implementation include a lack of cross-region evaluation, higher computational cost, and inconsistent explanations.

3. Proposed Methodology

Figure 1 shows the proposed methodology of multimodal and explainable agriculture sustainability.

The inputs are taken with different modalities including RGB images, thermal images, 3D point cloud and time series data. These inputs are pre-processed as per the requirement and passed to the model, Physio-Aware Multimodal Transformer network (PAMT-Net). This model consists of the different transformers for different modalities. These include, vision transformer, physics-informed fusion which uses cross-model attention, and temporal transformer. It also considers the physics constraint loss. Different performance measures are considered for the evaluation of the proposed models such as precision, recall, accuracy, mean absolute error. In addition to these the explainability and interpretability are also to be considered for the evaluation of trust in the diagnosis and analysis. Table 1 shows the feature extractions as per the modality.

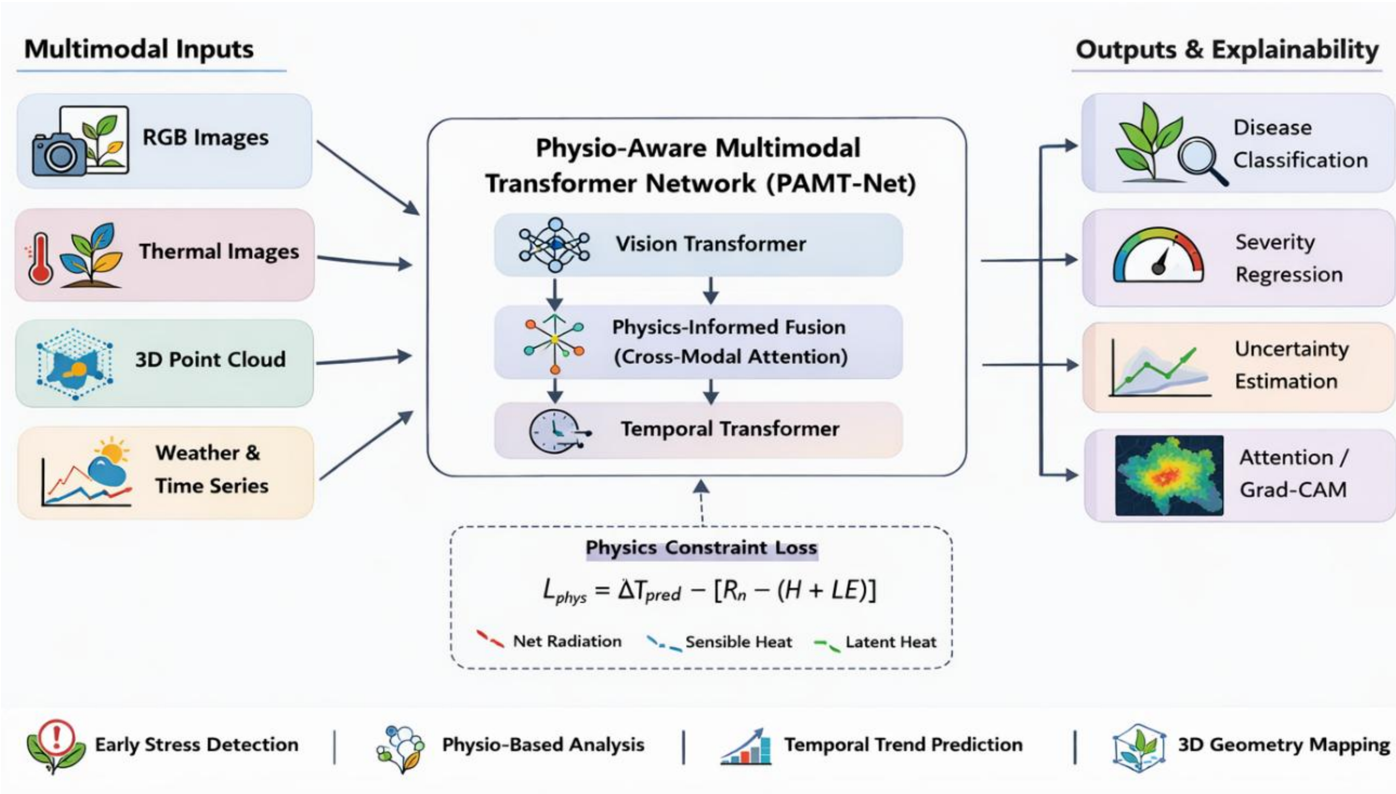

Figure 1: Proposed Methodology

Table 1: Feature Extraction

Modality	Model
Image	CNN / Vision Transformer
Time-series	LSTM / Transformer encoder
Text	BERT-like encoder

Explainability consists of the frameworks, namely intrinsic attention maps [4], SHAP [5] for tabular sensors, and Grad-CAM [6] for images. Uncertainty could be estimated by using Monte-Carlo dropouts [7].

4. Dataset

Table 2 shows the description of the dataset to be used for the implementation. Pheno4D dataset [8] is to be utilized for the model implementation and evaluation of the performance of the model. The dataset consists of the species including tomato and maize. It also consists of the 3Dscans and leaf-level annotations.

Table 2: Dataset description

Category	Description
Dataset Name	Pheno4D
Developed By	Technical University of Munich (TUM) & University of Bonn
Primary Focus	4D Plant Phenotyping
Dimension	3D Geometry + Temporal (4D)
Environment	Controlled Greenhouse
Plant Species	Tomato & Maize
Total Plants	~7 Plants (original benchmark release)
Total Time Points	~7–8 time points per plant
Total 3D Scans	~50–60 full plant scans
Leaf-Level Annotations	~1,500+ annotated leaves

Table 3 explains the different modalities of the dataset. It consists of 3Dpoint cloud, RGB images, temporal sequences, leaf-level labels and environmental metadata.

Table 3. Modalities of data

Modality	Data Type	Sample Volume	Purpose
3D Point Clouds	High-resolution laser scans	~50–60 scans	Structural modeling
RGB Images	Visual color images	Paired with scans	Texture & visual cues
Temporal Sequences	Repeated scanning over days	~7–8 time points/plant	Growth tracking
Leaf-Level Labels	Manual segmentation	~1,500+ leaves	Instance segmentation
Environmental Metadata	Greenhouse conditions	Experimental metadata	Controlled benchmarking

5. Sustainability Impact

The proposed work is aligned with the following SDGs and provide the following impact.

- SDG 2 (Zero Hunger)
- SDG 12 (Responsible Consumption)
- SDG 13 (Climate Action)

The proposed work will provide improvements like water reduction 10–20%, fertilizer optimization, and reduced false pesticide usage.

6. Conclusion

The proposed MM-XAI framework integrates multimodal fusion, attention-based explainability, and uncertainty modelling for sustainable agriculture. The architecture addresses limitations of existing systems by incorporating adaptive fusion, intrinsic explainability, and deployment-aware optimization. Expected outcomes indicate significant improvements in predictive performance, trust, and sustainability metrics. Future work includes large-scale field validation and farmer-cantered usability studies.

References

1. 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.
2. Shams, M. Y., Gamel, S. A., & Talaat, F. M. (2024). Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making. *Neural Computing and Applications*, *36*(11), 5695-5714.
3. Wang, S., Wang, C., Zhang, C., Xue, J., Wang, P., Wang, X., ... & Huo, Z. (2022). A classification-based spatiotemporal adaptive fusion model for the evaluation of remotely sensed evapotranspiration in heterogeneous irrigated agricultural area. *Remote Sensing of Environment*, *273*, 112962.
4. Gao, Y., Zeng, Z., Du, D., Cao, S., Zhou, P., Qi, J., ... & Yang, M. (2024). Seerattention: Learning intrinsic sparse attention in your llms. *arXiv preprint arXiv:2410.13276*.
5. Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022, October). SHAP-based explanation methods: a review for NLP interpretability. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4593-4603).
6. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, *128*(2), 336-359.
7. Milanés-Hermosilla, D., Trujillo Codorniú, R., López-Baracaldo, R., Sagaró-Zamora, R., Delisle-Rodriguez, D., Villarejo-Mayor, J. J., & Nunez-Alvarez, J. R. (2021). Monte carlo dropout for uncertainty estimation and motor imagery classification. *Sensors*, *21*(21), 7241.
8. Schunck, D., Magistri, F., Rosu, R. A., Cornelißen, A., Chebrolu, N., Paulus, S., ... & Klingbeil, L. (2021). Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *Plos one*, *16*(8), e0256340.