WORLD CONFERENCE ON STUDENT - CENTERED LEARNING APPROACHES

A COMPARATIVE ANALYSIS OF UNIMODAL AND MULTIMODAL DEEP LEARNING METHODS FOR THE EARLY DETECTION OF CROP DISEASES

Sindarov O.M.

Samarkand State University named after Sharof Rashidov

Abstract. The thesis presents a comparative analysis of unimodal (image-only) and multimodal (image, spectral, sensor, and meteorological data based) deep learning approaches for the early detection of crop diseases. The advantages and limitations of both directions are systematized by criteria, and the superiority of the multimodal approach under field conditions is substantiated.

Keywords: *artificial intelligence, deep learning, convolutional neural network, multimodal data, crop diseases, early diagnosis, data fusion.*

Introduction

The incidence of disease in agricultural crops is causing a substantial worldwide decline in yields and is increasingly threatening food security. Early-stage detection of disease enables the rational use of chemical agents, prevents yield losses, and allows timely agrotechnical decisions. Over the past decade, artificial intelligence — and deep learning algorithms in particular — has been widely adopted for the automated detection of crop diseases; however, the methods in use fall into two fundamentally different directions: unimodal methods that rely solely on image data, and multimodal methods that combine imagery with other types of data.

The purpose of this thesis is to carry out a comparative analysis of the operating principles, application domains, advantages, and limitations of both directions based on clearly defined criteria, and to substantiate why the multimodal approach may prove preferable under real field conditions.

Analysis of unimodal (image-only) methods

The most widespread approach to detecting crop diseases is the classification of leaf or stem images by disease type using convolutional neural networks (CNNs). Architectures such as ResNet, EfficientNet, and MobileNet achieve high accuracy on datasets collected under controlled, uniform lighting and background conditions (for example, laboratory-based datasets such as PlantVillage) [1; 7]. Such results demonstrate the strong capacity of these model architectures to extract visual features, yet this accuracy is often not directly transferable to real field conditions.

Object-detection models from the YOLO family enable real-time localization of diseased regions within a frame, a feature of considerable importance for practical monitoring systems. However, like classification-based CNNs, these models also rely solely on visual cues — changes in color, shape, and texture. As a result, accuracy declines sharply when illumination intensity, camera angle, or background complexity changes [3]. Moreover, such models are entirely unable to detect the stage of disease that has already begun physiologically but has not yet affected outward appearance, since their input consists of image data alone.

In recent studies, attention-based architectures — including the Vision Transformer — are being considered as an alternative to CNNs. They capture the global context of an image more effectively, but require large volumes of labeled data and substantial computational resources. A decline in accuracy when a model is generalized to a new crop variety or a different geographic region (the generalization problem) has been noted as a common shortcoming of all unimodal methods [3; 7].

Analysis of multimodal methods

Multimodal approaches analyze image data jointly with other sources — spectral imagery, temperature, humidity, and soil indicators obtained from IoT sensors, as well as meteorological data. Such hybrid models demonstrate greater robustness than unimodal approaches, since they rely not only on the external (visual) signs of disease but also on the environmental factors that trigger or accelerate it [2; 4].

Three strategies are distinguished according to the level at which data fusion takes place: early fusion — combination at the level of raw data; late fusion — combination of results after each modality has been classified separately; and intermediate fusion — combination at the level of feature vectors, after each modality has been transformed by its own encoder. Comparative studies show that intermediate fusion typically yields the highest accuracy, since it allows the interdependence between modalities to be captured more deeply [4].

In application-oriented research, hybrid frameworks have also been proposed that combine disease detection with treatment recommendation and even market-price forecasting — turning such systems into a comprehensive decision-support tool for farmers [5]. At the same time, foundation models pre-trained on large volumes of data, when combined with multimodal deep learning, are regarded as a promising direction in precision agriculture: they demonstrate strong generalization even with limited labeled data, although their high computational requirements complicate deployment on resource-constrained (edge) devices [4; 6].

Comparative analysis of the methods

To clearly illustrate the key differences between the two directions reviewed above, a comparative table is presented below (Table 1). The table compares each type of approach in terms of data source, principal advantage, and principal limitation.

Criterion	Unimodal (image)	Multimodal	Note / source
Data source	RGB/spectral image only	Image + sensor + meteorological data	[1; 2; 4]
Early-stage detection	Limited — only once a visual sign appears	Broader — physiological factors are taken into account	[2; 4]
Robustness to field	Low — degrades	High — multiple	[2; 3; 4]

conditions	under lighting/background influence	sources compensate for error	
Computational complexity	Relatively low	High — multiple encoders required	[4; 6]
Suitability for edge devices	Good (compact models available)	Requires additional optimization	[6; 7]
Practical readiness level	High — many ready-made solutions exist	Moderate — mostly at the research stage	[3; 5]

Table 1. Comparative analysis of unimodal and multimodal approaches

Discussion

The analysis presented shows that although unimodal methods are computationally simple and sufficiently mature for implementation, their principal weakness lies in their robustness being highly dependent on environmental conditions and in their inability to detect the early, visually imperceptible stage of disease. Multimodal methods are specifically aimed at overcoming these two limitations: sensor and meteorological data reflect factors that indirectly affect plant physiology, providing a warning signal even at a stage when the disease has not yet produced any external sign.

At the same time, the practical implementation of the multimodal approach is associated with certain difficulties: the scarcity of open datasets that jointly contain image and sensor data, the cost of deploying IoT infrastructure, and the need to compress the model for deployment on resource-constrained devices. Addressing these challenges constitutes the subject of the subsequent stages of this research.

Characteristics by crop type and disease group

The effectiveness of unimodal and multimodal methods also varies depending on the crop type and the disease group. Diseases of fungal origin (such as late blight, gray mold, and blight) typically produce a relatively early and clearly visible sign on the leaf surface (spots, discoloration), and in such cases unimodal CNN models may also achieve satisfactory results. However, viral and bacterial diseases, as well as abiotic stress conditions arising from water or nutrient deficiency, often manifest first not through visible signs but through the plant's physiological indicators — for example, leaf temperature or transpiration intensity. It is precisely in such cases that combining sensor data with imagery becomes decisive, since image-only analysis alone cannot detect such diseases until a late stage.

From this standpoint, the advantage of the multimodal approach is not uniform across all disease types but varies depending on the biological nature of the disease. This conclusion indicates that, when constructing the dataset for the proposed research, diseases of different origins must be covered separately, since only in this way can the model deliver reliable results across all disease groups.

Practical recommendations and directions for further research

Based on the analysis conducted, the following practical recommendation can be formulated: rather than relying on a single modality, it is advisable to begin the development of new diagnostic systems with a minimal multimodal scheme that combines imagery with at least one additional sensor indicator (such as leaf temperature or humidity), since this significantly improves accuracy without substantially increasing computational load. Additional modalities (meteorological data, soil analysis) can then be incorporated progressively, depending on available resources and data.

The next stage of the research envisages enriching the theoretical analysis presented in this thesis with experimental verification: jointly collecting image and sensor data under real field conditions, training the proposed fusion architecture, and

quantitatively comparing it against existing unimodal baseline models. Such experimental results will provide practical confirmation of the theoretical conclusions presented in this thesis and will reinforce the scientific novelty of the research.

Conclusion

The comparative analysis conducted confirms the theoretical superiority of multimodal deep learning methods over unimodal methods for the early detection of crop diseases, particularly with respect to robustness under field conditions and early-diagnosis criteria. At the same time, the practical readiness level of multimodal solutions currently remains lower than that of unimodal solutions, a circumstance that makes the development of new architectures and datasets in this direction a pressing scientific and practical task.

References

1. Dolatabadian A. [et al.]. Image-based crop disease detection using machine learning // Plant Pathology. – 2025. – DOI: 10.1111/ppa.14006.
2. Hamid [et al.]. Enhancing Agricultural Disease Detection: A Multi-Model Deep Learning Novel Approach // Engineering Reports. – 2025. – DOI: 10.1002/eng2.13113.
3. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture // Artificial Intelligence Review. – 2024. – DOI: 10.1007/s10462-024-11100-x.
4. Multimodal Deep Learning and Foundation Models for Early Detection and Forecasting of Plant Diseases // Plants. – 2026. – Vol. 15, No. 17. – Art. 2564. – DOI: 10.3390/plants15172564.
5. Sable N.P., Shukla V., Khedkar V., Ujalambkar D.M. Hybrid multimodal learning framework for crop disease detection, adaptive treatment, and price forecasting // Frontiers in Computer Science. – 2026. – DOI: 10.3389/fcomp.2026.1855851.

6. Artificial Intelligence-Driven Plant Disease Detection and Diagnosis: A Comprehensive Review of Deep Learning Approaches, Multimodal Sensing Technologies, and Future Perspectives in Precision Agriculture // The Plant Pathology Journal. – 2026. – DOI: 10.5423/PPJ.RW.01.2026.0004.
7. A Review on the Detection of Plant Disease Using Machine Learning and Deep Learning Approaches // Journal of Imaging. – 2025. – DOI: 10.3390/jimaging11100326.