A Review of Multimodal Fusion Techniques in Crop Disease Detection

Chaitanya Yadav [23FE10CSE00409]
Department of Computer Science
Manipal University Jaipur, Jaipur 303007, Rajasthan , India
Email: CHAITANYA.23FE10CSE00409@MUJ.MANIPAL.EDU

Abstract

Yield of crops with time has benefited from the progress and advancements in deep learning but approaches relying on single modal-data usually find it challenging to confidently capture the complex and diverse characteristics of plant-pathogen interactions. In recent years, multimodal learning which incorporates information from multiple data sources such as visual , thermal , spectral , sensor based data etc has come up as a highly promising solution for better detection accuracy and robustness.the following review presents a systematic overview of multimodal datasets and fusion strategies applied to plant disease detection.Various fusion approaches are analyzed and compared on the basis of performance , practical applicability and scalability. In addition to this the paper also discusses the shortcoming of deep learning approaches currently used in disease detection in plants and the need for multimodal approaches and the challenges associated with the same.By summarizing current research trends , this review aims to highlight the potential of multimodal fusion techniques and try to identify safest future research direction.

Keywords

Plant Disease Detection, Multimodal Learning, Multimodal Data Fusion, Deep Learning, Agriculture Image Analysis, Multimodal Datasets , Precision Agriculture

1 Introduction

1.1 Traditional Manual Detection and Its Challenges

Agriculture is a fundamental pillar of the global economy, providing livelihoods for billions and meeting the essential food needs of humanity [4, 5]. However, plant diseases significantly threaten this sector, causing global food production losses estimated between 20% and 40% [1, 2]. Historically, the identification of these diseases has relied on manual and visual inspection. Farmers and experts typically assess observable plant characteristics—phenotypes—such as leaf color, lesions, or blight spots with the naked eye to detect signs of infection [2].

While manual inspection remains common in many regions, it is inherently labor-intensive, time-consuming, and difficult to scale for large agricultural plantations [4, 5]. Furthermore, this method is prone to human error and often results in late-stage detection, after significant damage has already occurred, which limits the effectiveness of subsequent control measures [2, 5].

1.2 The Shift to Machine Learning and Deep Learning

To address the limitations of manual methods, there has been a significant shift toward automated solutions leveraging Artificial Intelligence (AI). Initially, machine learning (ML) algorithms like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF) were employed to automate pattern and anomaly detection [3, 6]. These methods provided a more objective and rapid alternative to manual surveys [2].

The field has since been revolutionized by Deep Learning (DL), which offers vastly superior performance to conventional methods by automatically extracting complex features from raw data [3]. Convolutional Neural Networks (CNNs) have emerged as the dominant architecture for image-based disease detection due to their high accuracy in classification and segmentation tasks [2, 4]. By mimicking the functional capabilities of the human brain, these models can identify subtle visual cues of pathogen infection that are often invisible to the naked eye [2, 3].

1.3 Limitations and Existing Approach

Despite these advancements, several challenges persist in current DL-based systems. Many models suffer from poor generalization because they are trained on datasets captured under controlled laboratory conditions with plain backgrounds [1, 5]. When deployed in real-world "farm-field" environments, their accuracy often drops due to complex backgrounds, varying lighting, and overlapping disease spots [1, 2].

Additionally, most existing research focuses solely on RGB imagery, which only captures symptoms visible in the light spectrum [2, 6]. This reliance makes it difficult to detect diseases at the earliest ”pre-symptomatic” stages, leading to a lack of farm-field practicality for timely intervention [2].

1.4 The Necessity of Multimodal Fusion

Multimodal fusion addresses these limitations by integrating diverse data types to provide a comprehensive view of plant health [7]. Instead of relying on a single source, these approaches combine various inputs such as RGB, multispectral, hyperspectral, thermal, and radar data [7].

The utility of multimodal fusion lies in its ability to detect physiological changes that are invisible to both the human eye and standard cameras [6, 7]. For example, thermal and hyperspectral sensors can identify early-stage pathogen contagions by detecting temperature shifts or nutrient changes before visual lesions appear [6]. In the context of Agriculture 5.0, fusing data from ground-based sensors, UAVs, and satellites allows for real-time, large-scale monitoring that is robust against environmental variability [7]. This holistic approach is essential for meeting the agricultural demands of climate change and ensuring global food security [7].

2 General Principles of Multimodal Learning

2.1 The Multimodal Paradigm and Objectives

The fundamental premise of multimodal learning is that our experience of the world is inherently multisensory. In computational terms, a modality refers to a specific type of data or the way in which information is encoded, such as images, text, audio, or sensory signals. Multimodal machine learning aims to build models that can process, relate, and consolidate information from these diverse sources to achieve a more robust understanding than is possible with a single modality [9]. The core objective of deep multimodal representation learning is to bridge the ”heterogeneity gap,” which refers to the different statistical properties and structures inherent in different data types, such as the grid-like structure of pixel data versus the sequential nature of text [10].

2.2 Core Technical Challenges

To effectively integrate multiple modalities, researchers must address five primary technical challenges:

Representation: Learning how to summarize multimodal data in a way that exploits the complementarity of the sources while maintaining their unique signals [9].

Translation: Mapping data from one modality to another, such as generating a descriptive text based on a visual input [9].

Alignment: Identifying the direct relationships between sub-elements of different modalities (e.g., matching a specific lesion in an image with a specific symptom mentioned in a report) [9].

Fusion: Integrating information from two or more modalities to perform a prediction task, such as classification or detection [9].

Co-learning: Transferring knowledge between modalities, particularly when one modality is "resource-poor" or limited in data compared to another [9].

2.3 Deep Representation Learning Frameworks

Deep learning has significantly advanced the ability to learn shared features across modalities without extensive hand-engineering [8]. These architectures generally fall into three structural categories:

Joint Representations: These models project unimodal features into a shared hidden space. This is often achieved through deep autoencoders or restricted Boltzmann machines that fuse the data into a single, unified feature vector used for the final task [8].

Coordinated Representations: Rather than merging modalities into one vector, coordinated frameworks learn separate but constrained representations. For example, the model ensures that the feature vector for an image and its corresponding text description are close together in a mathematical space (similarity-based) or highly correlated (correlation-based) [10].

Cross-Modality Feature Learning: This approach demonstrates that learning features for one modality can be significantly improved if other modalities are present during the training phase, even if those extra modalities are absent during the actual testing phase [8].

3 Multimodal Data Integration Strategies

The integration of diverse data modalities has become a cornerstone of precision agriculture, enabling models to overcome the limitations of individual sensors by capturing a more holistic view of crop health and environmental conditions. Recent research highlights four primary fusion paradigms: image and weather data, image and satellite data, image and temporal sequences, and image and soil sensor inputs.

Table 1: Comparison of Multimodal Fusion Architectures

Study	**Primary Modalities**	**Model Architecture**	**Key Features**
Maillet et al. (2025)	Satellite + Weather	ViT + Transformer Encoders	Bottleneck fusion mode; missing image interpolation
Chityala (2022)	Satellite + Weather + Soil	ResNet-50 + Bi-LSTM + FFN	Cross-attention fusion mechanism
Trupthi et al. (2025)	Image + Soil/Temp/Hum	CNN-Transformer + SBOA	Retrieval-Augmented Generation (RAG) module
Godara (2025)	Image + IoT Sensors	EfficientNetB0 + Dense Layers	Real-time sensor processing with Grad-CAM
Masrur et al. (2024)	UAS + Satellite	SRCNN	Spectral and spatial extension for biomass

3.1 Image and Weather Fusion

The synergy between visual plant characteristics and meteorological conditions is critical for detecting diseases like downy mildew, which are heavily influenced by environmental variables such as humidity and temperature [11].

Architectures: Modern approaches utilize transformer-based models to handle the heterogeneity of 2D visual features and 1D weather sequences. For instance, a three-component architecture using a Vision Transformer (ViT) for image features and two transformer-encoders for weather data achieved 97% accuracy in disease detection [11].

Frameworks: Other frameworks, such as AgroFusionNet, employ a multi-branch system where ResNet-50 extracts spatial features from images while a stacked Bi-LSTM branch processes temporal weather data (precipitation, radiation, temperature) to capture climate stress events [12].

IoT Integration: IoT-enabled systems have also been developed to combine real-time sensor streams—including rainfall and wind speed—with EfficientNet-based leaf analysis to provide context-aware diagnostics [14]. These systems are designed to maintain reliability even when sensor readings include Gaussian noise [14].

3.2 Image and Satellite Fusion

Fusing high-resolution Unmanned Aircraft System (UAS) imagery with broad-coverage satellite data addresses the trade-off between spatial detail and scalability.

Super-Resolution Techniques: Frameworks like the ”spectral extension” model use neural network-based super-resolution (SRCNN) to enhance 10m Sentinel-2 bands to sub-meter resolution by leveraging UAS RGB data [15]. This allows the reconstruction of critical vegetation indices, such as Red Edge (VRE) and Near-Infrared (NIR), which are typically missing from standard RGB cameras [15].

Applications: These fusion models have improved the accuracy of biomass and nitrogen estimation by 18% and 31%, respectively, compared to using 10m satellite data alone [5]. Additionally, 3D convolutional attention modules have been proposed to align multitemporal data for improved crop classification and identification [18].

3.3 Image and Temporal Fusion

Temporal fusion tracks crop health over the growing season, capturing the progression of diseases and growth stages [16].

Missing Data Interpolation: Because satellite imagery can be hindered by cloud cover, ConvLSTM models are increasingly used to interpolate missing intermediate images, creating a daily temporal dataset for continuous monitoring [11].

Spatio-Temporal Modeling: Advanced models integrate recurrent 3D convolutional neural networks to simultaneously process spatial and temporal dimensions of imagery [12][12]. These models analyze variables such as crop growth stages and weather fluctuations over time to provide dynamic assessments of yield potential [16].

3.4 Image and Soil Sensor Fusion

Combining leaf imagery with subsurface soil conditions allows for early disease detection, often before visual symptoms appear [17].

Fusion Paradigms: Research methodologies such as RAG-MMF-SF utilize CNN-based Transformers to fuse plant imagery with time-synchronized soil moisture, pH, and nutrient sensors [13].

Optimization and Retrieval: To enhance decision support, some systems incorporate Retrieval-Augmented Generation (RAG) to cross-reference fused sensor-image data with external agronomic knowledge bases [13].

Performance: Studies integrating static soil attributes through fully connected networks alongside image branches have demonstrated significant accuracy gains, achieving up to 97.54% accuracy in multi-modal configurations compared to unimodal baselines [13]. Furthermore, multi-modal sensor input assemblies are being developed to determine crop diseases at various phases of growth [17].

Table 2: Performance Comparison of Multimodal Models in Agricultural Applications

Model/Method	**Task**	**Best Result**	**Data Split**
RAG-MMF-SF	Crop Health Diagnosis	97.54% Accuracy	80–20 Split
ViT-Weather Fusion	Downy Mildew Detection	97.0% Accuracy	Multi-year test
AgriDeepFusionNet	Multi-stage Disease Detection	96.8% Accuracy	–
MCYP-Net	Crop Yield Prediction	0.91 R^2	Multi-year datasets
SRCNN-Fusion	Nitrogen Estimation	31% Improvement	–
IoT-Enabled DL	Large-scale Disease Pred.	93.8% Accuracy	PlantVillage (3000 classes)

4 Constraints and Limitations of Fusion Paradigms

Despite the performance gains offered by multimodal architectures, each data combination presents unique technical and operational hurdles that must be managed to ensure reliable field deployment.

4.1 Image and Weather Fusion

Heterogeneity Gap: Merging 2D visual feature maps with 1D sequential weather data remains a challenge in architecture design, often requiring complex transformer-based ”bottleneck” tokens to align the different data types [11].

Station Proximity: The effectiveness of these models is highly dependent on the location of meteorological stations; localized microclimates may not be accurately captured if sensors are distant from the specific plot being analyzed [12].

Visual Gaps: While weather data is generally continuous, the accompanying satellite imagery is frequently obstructed by cloud cover, necessitating sophisticated interpolation models like ConvLSTM to fill temporal gaps [11].

4.2 Image + Satellite Fusion

Resolution Disparity: The significant scale difference between high-resolution UAS imagery (cm-scale) and satellite data (m-scale) makes pixel-perfect alignment difficult, potentially leading to registration errors [15].

Computational Overhead: Enhancing satellite bands through super-resolution (SRCNN) requires high-end GPU resources, which can be a barrier for real-time agricultural applications [15].

Spectral Complexity: Managing high-dimensional data across different satellite constellations while maintaining spectral consistency for crop classification is still an evolving challenge [18].

4.3 Image and Temporal Fusion

Interpolation Risks: Filling gaps in temporal sequences can introduce ”hallucinated” data that may not reflect sudden, rapid-onset disease outbreaks [11].

Computational Intensity: Analyzing long-term spatio-temporal sequences using 3D CNNs or LSTMs requires massive memory and processing power, making it difficult to deploy on lightweight edge devices [12][16].

4.4 Image and Soil Sensor Fusion

Sensor Reliability: In-situ sensors are vulnerable to environmental degradation, leading to signal drift or Gaussian noise which can negatively impact diagnostic accuracy [14].

Scalability Constraints: Unlike remote sensing, soil sensors provide point-specific data. Generalizing this information across large, diverse landscapes requires a dense, cost-prohibitive sensor network [13][1].

Knowledge Integration: Integrating static soil properties with dynamic plant imagery requires advanced retrieval mechanisms (like RAG) to provide meaningful context, adding layers of complexity to the AI framework [13].

5 Conclusion and Future Direction

This study has compared four distinct multimodal fusion approaches, each demonstrating clear advantages over unimodal systems. However, based on the comparative analysis of accuracy, scalability, and implementation cost, Image + Weather fusion stands out as the most robust and promising ”safest bet” for future research [11][12][14].

The superiority of this paradigm is driven by three primary factors:

Diagnostic Synergy: Weather variables (such as humidity and temperature) provide the environmental ”drivers” of disease, while imagery provides the ”physical evidence.” Combining these allows models to achieve accuracies as high as 97% [11].

Infrastructure Accessibility: Satellite imagery (e.g., Sentinel-2) and meteorological data are often freely available or already part of existing farm infrastructure, making this fusion more accessible than high-cost UAS flights or dense soil sensor grids [15].

Global Scalability: Because weather and satellite data can be gathered remotely via global grids, this approach can be scaled across varied agro-climatic zones without the need for site-specific hardware installation [14].

Future research should focus on refining the real-time processing of these fused streams and exploring the use of Retrieval-Augmented Generation (RAG) to translate these complex AI outputs into actionable, plain-language advice for farmers on the ground [13]. Improving the robustness of these models against sensor noise and cloud interference will be essential for the next generation of smart farming diagnostic tools [14]17.

References

[1] K. Antwi, K. E. Bennin, D. K. P. Asiedu, and B. Tekinerdogan, ”On the application of image augmentation for plant disease detection: A systematic literature review,” Smart Agricultural Technology, vol. 9, 2024.

[2] A. Upadhyay et al., ”Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture,” Artificial Intelligence Review, vol. 58, no. 92, 2025.

[3] H. N. Ngugi, A. E. Ezugwu, A. A. Akinyelu, and L. Abualigah, ”Revolutionizing crop disease detection with computational deep learning: a comprehensive review,” Environmental Monitoring and Assessment, vol. 196, no. 302, 2024.

[4] I. Pacal et al., ”A systematic review of deep learning techniques for plant diseases,” Artificial Intelligence Review, vol. 57, no. 304, 2024.

[5] P. Sajitha, A. D. Andrushia, N. Anand, and M. Z. Naser, ”A review on machine learning and deep learning image-based plant disease classification for industrial farming systems,” Journal of Industrial Information Integration, vol. 38, 2024..

[6] A. Dolatabadian et al., ”Image-based crop disease detection using machine learning,” Plant Pathology, vol. 74, no. 1, pp. 18–38, 2025.

[7] M. El Sakka, M. Ivanovici, L. Chaari, and J. Mothe, ”A Review of CNN Applications in Smart Agriculture Using Multimodal Data,” Sensors, vol. 25, no. 472, 2025.

[8] . Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal Deep Learning,” in Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019.

[10] W. Guo, J. Wang, and S. Wang, “Deep Multimodal Representation Learning: A Survey,” IEEE Access, vol. 7, pp. 64083–64094, 2019.

[11] W. Maillet, M. Ouhami, and A. Hafiane, "Fusion of Satellite Images and Weather Data with Transformer Networks for Downy Mildew Disease Detection," arXiv, 2025.

[12] S. Chityala, "AgroFusionNet: A multi-modal AI framework for predictive crop yield modeling using satellite imagery, weather patterns, and soil data," International Journal of Engineering in Computer Science, vol. 4, no. 2, pp. 67-74, 2022.

[13] M. Trupthi et al., "RAG-Enhanced Smart Farming: A Methodology for Multimodal Fusion and AI-driven Crop Health Diagnosis," International Journal of Intelligent Engineering and Systems, vol. 18, no. 10, pp. 655-668, 2025.

[14] B. Godara, "IOT-Enabled Deep Learning Framework for Scalable Crop Disease Prediction Using Multimodal Sensor and Image Fusion," IJIRT, vol. 12, no. 7, 2025.

[15] A. Masrur et al., "Learning to See More: UAS-Guided Super-Resolution of Satellite Imagery for Precision Agriculture," arXiv, 2024.

[16] T. M. Bhalodia and A. M. Kothari, "Hyperspectral Images and Temporal Data Fusion using Machine Learning for Crop Health Monitoring and Yield Prediction," Journal of Systems Engineering and Electronics, 2025.

[17] B. Verma, A. Kumar, D. Kumar, and R. K. Dwivedi, "AI-driven predictive modeling framework for early detection of multi-stage crop diseases using multi-modal sensor data and deep transfer learning approaches," IJAFS, vol. 7, no. 9, pp. 36-41, 2025.

[18] Z. Mo and Y. Wu, "Precision crop yield prediction model based on deep learning and multimodal data fusion," Turkish Journal of Agriculture and Forestry, vol. 49, no. 3, pp. 580-597, 2025.