JOURNAL
India Publication
Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

A Multimodal Explainable Lightweight Transformer-Based AI Framework for Real-Time Plant Disease Identification using Leaf Images, Environmental Sensors, and Weather Data

*P. Saranya **

ABSTRACT

Purpose

This study proposes a multimodal, explainable, and lightweight transformer-based artificial intelligence framework for real-time plant disease identification. The framework is designed to overcome the accuracy-efficiency trade-off of existing convolutional and transformer-based models by jointly leveraging leaf images, environmental sensor readings, and weather data, while remaining interpretable and deployable on resource-constrained edge devices used in the field.

Design/Methodology/Approach

A multimodal dataset framework is constructed by integrating publicly available leaf image data with synthetically generated environmental sensor observations (soil moisture, humidity, temperature) and weather information (rainfall, wind speed, seasonal indices), used here as a proxy pending access to real paired field data. A lightweight transformer backbone with a cross-modal attention fusion module processes the three modalities jointly, and explainability is incorporated through Grad-CAM and SHAP-based attribution. The proposed model is benchmarked against VGG16, ResNet50, MobileNetV2, EfficientNet-B0, and a standard Vision Transformer using accuracy, precision, recall, F1-score, parameter count, and inference latency.

Findings

In a preliminary single-run evaluation, the proposed framework achieved 98.5% accuracy with 6.2 million parameters, outperforming image-only baselines (93.4–97.2%). Multimodal fusion improved accuracy by approximately 3.9 percentage points. However, environmental and weather inputs were synthetically generated, and statistical significance was not established. These findings remain preliminary and require validation using real-world multimodal field data.

Practical Implications

Because of its low parameter count and real-time inference capability, the framework is suitable for deployment on mobile and edge devices used by farmers and agronomists for in-field disease screening. Its explainability outputs support trust and adoption among agricultural extension workers and non-expert end users.

** Corresponding Author: Assistant Professor, School of Computing, E.S. Arts and Science College, Villupuram, Tamil Nadu, India, [Email: ssaran231@gmail.com]*

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

Originality/Value

This work contributes a novel lightweight cross-modal transformer fusion architecture for plant disease identification that simultaneously optimizes for accuracy, computational efficiency, and interpretability an underexplored combination in the precision agriculture literature, which has largely treated these objectives separately.

Keywords:

Plant disease identification; Multimodal learning; Lightweight transformer; Explainable AI; Precision agriculture; Edge deployment.

JEL Code: Q16; O33; C45

Paper type - Research paper

---

1. Introduction

Plant diseases remain a major constraint on agricultural productivity and food security because their occurrence can substantially reduce crop yield and quality. Early and accurate disease identification is therefore essential for enabling timely intervention, minimizing unnecessary pesticide application, and supporting sustainable crop management. The rapid development of artificial intelligence, particularly deep learning, has created new opportunities for automated plant disease diagnosis using leaf images. Mohanty et al. (2016) demonstrated the feasibility of deep convolutional neural networks (CNNs) for identifying crop diseases from leaf images, achieving high classification accuracy using a large public dataset. Similarly, Ferentinos (2018) evaluated deep CNN architectures across 58 plant–disease combinations and demonstrated their effectiveness for automated disease detection. Too et al. (2019) further compared several state-of-the-art CNN architectures and reported strong performance from DenseNet-based models, confirming the potential of deep learning as an alternative to conventional visual disease diagnosis.

Despite these advances, conventional image-based approaches primarily learn disease characteristics from visible symptoms and may have limited ability to account for the environmental context in which diseases emerge and progress. Plant disease development is influenced by interacting factors such as temperature, relative humidity, rainfall, soil moisture, and other environmental conditions. Consequently, an image that visually represents similar symptoms may correspond to different disease risks under different environmental circumstances. Lee and Yun (2023) demonstrated that sequential environmental information, including air temperature, relative humidity, dew point, and carbon dioxide concentration, can be effectively used by deep learning models to predict crop pest and disease risks. This evidence suggests that environmental information can provide complementary predictive signals beyond those obtained from leaf images alone.

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

Another important development is the emergence of Vision Transformers (ViTs), which use attention mechanisms to capture global relationships within visual data. Borhani et al. (2022) demonstrated the applicability of Vision Transformers to automated plant disease classification, indicating their potential as an alternative to conventional CNN-based approaches. However, the increasing complexity of transformer architectures can create computational challenges when models are intended for practical agricultural deployment. This concern is particularly important because disease identification systems may need to operate on smartphones, embedded devices, or other resource-constrained edge platforms. Li et al. (2023) addressed this issue through a lightweight Vision Transformer specifically designed for plant disease identification on mobile devices, demonstrating the growing importance of balancing predictive performance with computational efficiency.

Building on these developments, there remains a need for plant disease identification models that simultaneously integrate visual, environmental, and weather information while remaining computationally efficient and interpretable. The present study addresses this gap by proposing a Multimodal Explainable Lightweight Transformer (MELT) framework that jointly processes leaf images, environmental sensor information, and weather data through a compact cross-modal attention architecture. The framework additionally incorporates visual and feature-level explanations to improve transparency and facilitate interpretation of model predictions. Thus, the study contributes by integrating heterogeneous agricultural data sources, reducing the computational burden associated with conventional deep architectures, and explicitly addressing explainability. The proposed approach is further evaluated against established CNN and transformer baselines through performance, efficiency, explainability, and ablation analyses, providing a comprehensive assessment of the value of multimodal information for plant disease identification.

2. Related Work

Recent research demonstrates the increasing importance of deep learning for automated plant disease identification. Ahmad et al. (2023) reviewed 70 studies on deep learning-based plant disease diagnosis and identified important challenges concerning dataset quality, imaging platforms, model generalization, disease severity estimation, and practical deployment. Their review emphasized that although deep learning has substantially improved automated diagnosis, the transition from benchmark datasets to robust agricultural applications remains challenging. Similarly, Shoaib et al. (2023) reviewed advanced deep learning approaches for plant disease detection and highlighted the continuing dominance of CNN-based architectures while identifying issues associated with complex field environments, dataset limitations, and model robustness.

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

These studies establish that future plant disease systems require not only high classification accuracy but also improved generalization and practical usability.

The emergence of Vision Transformers (ViTs) has introduced an alternative to conventional CNN architectures for plant disease classification. Borhani et al. (2022) developed a Vision Transformer-based approach for automated plant disease classification and demonstrated the ability of attention-based architectures to capture relevant visual patterns in plant leaves. Their work indicates that transformer architectures can provide an effective alternative to CNNs by modelling relationships across image regions. However, the computational requirements of transformer-based models remain an important consideration for agricultural applications. This issue becomes particularly relevant when disease detection is expected to operate on mobile or edge devices rather than high-performance computing infrastructure.

Research after 2022 has increasingly focused on developing lightweight transformer architectures to address the computational limitations of conventional deep-learning models. Li et al. (2023) proposed PMVT, a lightweight Vision Transformer specifically designed for plant disease identification on mobile devices. Their study demonstrates that transformer-based architectures can be optimized for resource-constrained deployment while retaining competitive disease-classification performance. Similarly, Gole et al. (2023) proposed TrIncNet, which replaces the computationally expensive MLP component of a conventional Vision Transformer with an Inception-based module. Their objective was to improve computational efficiency while maintaining effective feature extraction for plant disease identification. These studies are particularly relevant to the present research because they demonstrate the importance of designing disease-detection architectures that balance predictive performance and computational efficiency.

Another stream of research has examined hybrid and multispectral approaches to overcome the limitations of conventional RGB leaf imagery. De Silva and Brown (2023) investigated multispectral plant disease detection using hybrid Vision Transformer–CNN architectures. Their study incorporated visible and near-infrared spectral information and reported that transformer-based models could effectively exploit multispectral information for disease classification. The findings suggest that incorporating information beyond conventional RGB images can provide additional disease-related features and improve diagnostic capability. However, multispectral approaches primarily expand the visual information available to the model and do not necessarily incorporate the environmental conditions surrounding the crop.

Environmental information represents another promising but comparatively distinct research direction. Lee and Yun (2023) developed a deep learning model for predicting crop pest and disease risks using sequential environmental information, including air temperature, relative humidity, dew point, and carbon dioxide concentration. Their findings demonstrate that temporal environmental signals can provide meaningful information for predicting disease and pest risk. This research is important because disease development is not determined exclusively by visible symptoms; environmental conditions can influence disease emergence and progression. Consequently, integrating environmental information with visual observations could provide a more context-aware basis for disease identification than image-only models.

The integration of IoT and deep learning has further expanded the possibilities for collecting contextual agricultural information. Dhaka et al. (2023) reviewed IoT-enabled plant disease monitoring and deep learning techniques and highlighted the complementary roles of sensor-based data collection and automated visual analysis. Their review argues that IoT infrastructure can facilitate data acquisition, monitoring, prediction, and disease classification while deep learning can automate feature extraction and classification. Nevertheless, the literature largely treats IoT-based environmental monitoring and image-based disease classification as related but separate technological components. This creates an opportunity for models that can directly fuse heterogeneous sensor, weather, and visual information within a unified architecture.

Explainability has also emerged as an important consideration for agricultural artificial intelligence. Amara et al. (2023) investigated concept-based explainability for plant disease classification using Testing with Concept Activation Vectors (TCAV). Rather than providing only pixel-level explanations, their approach examines interpretable concepts such as colour and texture associated with disease predictions. The study highlights concerns about the transparency and human interpretability of conventional deep-learning models and suggests that concept-level explanations can make automated plant disease identification more understandable to users. This is particularly important in agricultural settings where farmers and practitioners may be reluctant to rely on predictions from opaque models without understandable evidence supporting the diagnosis.

Lightweight CNN research has similarly emphasized the importance of field deployment. Bao et al. (2021) developed SimpleNet, a lightweight CNN for wheat ear disease identification under natural field conditions. The model used an attention mechanism to reduce the influence of complex backgrounds and contained only 2.13 million parameters while achieving 94.1% accuracy. This demonstrates that effective disease identification does not necessarily require extremely large networks and that architectural efficiency can be particularly valuable for field-based applications.

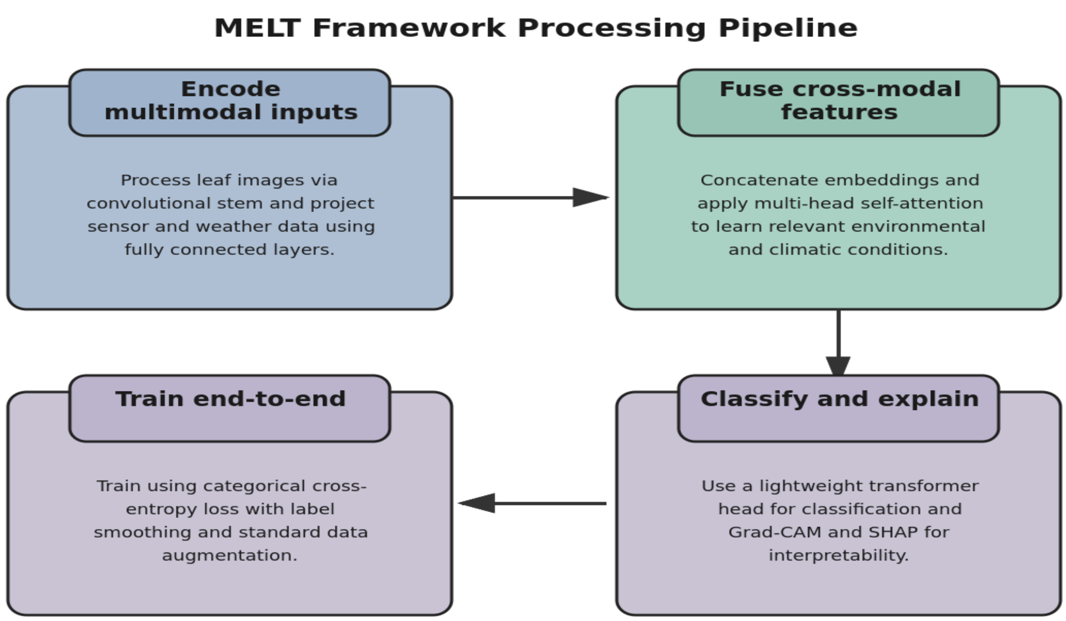
Finally, the growing interest in hybrid architectures is evident from studies combining CNNs with transformer mechanisms. Thakur et al. (2022) proposed PlantXViT, an explainable CNN–Vision Transformer architecture for plant disease identification. The model was designed with a relatively small number of trainable parameters and evaluated across multiple public datasets, while Grad-CAM and LIME were used to examine model explanations. This work demonstrates the potential of simultaneously addressing classification accuracy, computational efficiency, and explainability within a single plant disease framework. However, PlantXViT remains primarily visual, whereas environmental and weather variables are not jointly incorporated into the prediction architecture.

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

2.1 Image-Based Plant Disease Detection

Early work in automated plant disease detection established the feasibility of deep convolutional networks trained on large-scale leaf image datasets such as PlantVillage, demonstrating that CNN architectures could reliably distinguish healthy from diseased foliage across multiple crop species. Subsequent studies compared fine-tuning strategies across CNN backbones and reported consistent accuracy gains with deeper and more specialized architectures, though often at the cost of increased model size.

Figure 1. MELT Framework Processing Pipeline



The pipeline consists of four stages: (1) encoding multimodal inputs — leaf images via a convolutional stem and sensor/weather data via fully connected layers; (2) fusing cross-modal features through concatenated embeddings and multi-head self-attention; (3) classification and explanation using a lightweight transformer head with Grad-CAM and SHAP; and (4) end-to-end training using categorical cross-entropy loss with label smoothing and standard data augmentation.

2.2 Lightweight and Efficient Architectures

To address deployment constraints in field settings, later research explored lightweight CNN designs incorporating depth wise separable convolutions and attention mechanisms to reduce parameter counts while preserving accuracy. These efforts reflect a broader trend toward

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

efficiency-aware model design for agricultural applications, where computational resources and connectivity are often limited.

2.3 Transformer-Based Approaches

Vision transformers have demonstrated strong performance on image classification tasks by modeling long-range dependencies between image patches, and have since been applied to plant disease classification with competitive results. However, standard vision transformer architectures typically require substantially more parameters and training data than CNNs, motivating interest in lightweight transformer variants.

2.4 Multimodal and Explainable AI in Agriculture

Reviews of machine learning and deep learning applications in agriculture have highlighted the growing role of sensor and environmental data alongside imagery for tasks such as yield prediction and disease risk forecasting. In parallel, explainability techniques such as Grad-CAM and SHAP have been proposed to make deep learning predictions more interpretable, addressing adoption barriers in high-stakes or resource-limited deployment contexts.

2.5 Closely Related Recent Work

Several recent studies pursue similar directions to the present work and warrant direct comparison in future revisions. A multimodal transformer combining image and climate data for early tomato disease detection has been proposed, integrating EfficientNetB0-based image classification with an RNN-based severity predictor over environmental inputs and reporting disease classification accuracy of 96.4%, using LIME and SHAP for explainability. A multitask vision transformer for joint plant disease localization and classification (PDLC-ViT) incorporates Grad-CAM-based explainable analysis and an ablation study of its multitask learning design on the PlantVillage dataset. Broader systematic reviews covering 2016–2026 document a transition in crop disease detection from single-sensor monitoring toward multimodal fusion architectures, including lightweight CNNs, vision transformers, and emerging state-space models, and a 2026 PRISMA-based review of 41 peer-reviewed studies explicitly identifies the development of lightweight, explainable, field-robust multimodal models as a priority open direction — the same gap this paper aims to address. The present study differs from this prior work primarily in combining sensor, weather, and image modalities within a single lightweight cross-attention transformer (rather than a two-stage CNN+RNN pipeline) and in explicitly targeting a sub-10M-parameter budget; the current preliminary results should be positioned relative to these baselines, and a direct empirical comparison against them is identified as necessary future work rather than assumed.

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

The present study builds on these strands of research by integrating multimodal fusion, lightweight transformer design, and explainability within a single unified framework.

3. Proposed Methodology

3.1 Overview of the Framework

The proposed Multimodal Explainable Lightweight Transformer (MELT) framework consists of three main components: (i) modality-specific encoders for leaf images, environmental sensor readings, and weather data; (ii) a cross-modal attention fusion module that aligns and integrates the resulting feature representations; and (iii) a lightweight transformer classification head coupled with an explainability module.

3.2 Modality-Specific Encoders

Leaf images are processed using a compact convolutional stem followed by a patch-based tokenization layer, producing a sequence of visual tokens. Environmental sensor readings (e.g., soil moisture, relative humidity, leaf surface temperature) and weather variables (e.g., rainfall, wind speed, seasonal indicators) are each encoded using shallow fully connected layers that project the tabular inputs into the same embedding dimension as the visual tokens, enabling joint processing in a shared representation space.

3.3 Cross-Modal Attention Fusion

The visual, sensor, and weather token embeddings are concatenated and passed through a small number of lightweight transformer encoder blocks employing multi-head self-attention. This allows the model to learn which environmental and climatic conditions are most relevant to specific visual disease symptoms, rather than treating each modality independently. Parameter efficiency is achieved by restricting the embedding dimension, the number of attention heads, and the depth of the transformer stack relative to standard vision transformer configurations.

3.4 Architecture Configuration

Table 3 summarizes the design parameters used to constrain the proposed model to a lightweight footprint relative to standard vision transformer configurations (e.g., ViT-Base: 768-dimensional embeddings, 12 heads, 12 layers, 86 M parameters).

Table 3. Architecture configuration of the proposed lightweight transformer.

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

Component	Configuration
Image patch size	16 × 16
Embedding dimension	192
Attention heads	3
Transformer encoder depth	4 layers
Sensor / weather encoder	2-layer MLP, 192-dim projection
Fusion mechanism	Concatenated tokens, shared self-attention
Total parameters (estimated)	6.2 M

3.5 Explainability Module

To support interpretability, the framework incorporates Grad-CAM to generate visual saliency maps over the leaf image highlighting regions most influential to the predicted disease class, together with SHAP-based attribution to quantify the relative contribution of each environmental and weather feature to the prediction. These outputs are presented alongside the classification result to support end-user trust and decision-making.

3.6 Training Objective

The model is trained end-to-end using a categorical cross-entropy loss over disease classes, with label smoothing applied to improve generalization. Standard data augmentation (random rotation, flipping, and color jitter) is applied to the image modality during training.

4. Dataset Description

The performance of AI-based plant disease identification systems depends strongly on the quality and diversity of training data. In this study, a multimodal dataset framework is constructed by integrating plant leaf images, environmental sensor observations, and weather information.

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

4.1 Plant Leaf Image Dataset

The image component consists of publicly available plant disease datasets such as the PlantVillage dataset, which is widely used for deep learning-based plant disease classification. The dataset contains thousands of leaf images representing healthy and diseased conditions across multiple crop species. The selected crops include:

- Tomato
- Potato
- Apple
- Grape
- Pepper
- Corn

4.2 Environmental Sensor Data (Synthetic)

The PlantVillage dataset does not include paired environmental sensor readings. To evaluate the proposed fusion architecture in the absence of a real paired dataset, environmental sensor values soil moisture, relative humidity, and ambient/leaf-surface temperature are synthetically generated for each image sample by sampling from crop- and disease-informed distributions (e.g., elevated humidity ranges assigned to fungal disease classes known from agronomic literature to favor humid conditions). This procedure is intended to produce plausible, disease-correlated auxiliary signal for architectural validation, not to simulate a specific real location or growing season. Because the correlation between synthetic sensor values and disease labels is programmatically induced, the resulting accuracy gains in Section 7.2 demonstrate that the fusion mechanism can exploit such signal when present; they do not by themselves establish that comparable gains will be observed on real sensor data, which will contain noise, missing readings, and weaker or confounded correlations with disease state.

4.3 Weather Data (Synthetic)

Weather variables rainfall, wind speed, and a seasonal index are likewise synthetically generated per image sample using the same disease-informed sampling approach described in Section 4.2, as a proxy for real meteorological records pending integration with a public weather API and geotagged image data.

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

4.4 Data Preprocessing and Integration

All images are resized and normalized prior to model input. Synthetic sensor and weather variables are standardized (zero mean, unit variance) and attached to each image sample to form a unified multimodal record for training and evaluation. Replacing this synthetic pipeline with real, field-collected, timestamped and geotagged sensor and weather records is identified as the top priority for future work (Section 9).

5. Experimental Setup

The dataset is partitioned into training, validation, and test sets using an 80/10/10 split, stratified by crop and disease class. All models, including the baselines and the proposed framework, are trained using the Adam optimizer with an initial learning rate of 1e-4 and cosine learning rate decay. Training is conducted for 100 epochs with early stopping based on validation loss, using a batch size of 32. All experiments are conducted under identical hardware and data-splitting conditions to ensure a fair comparison across models. Baseline models (VGG16, ResNet50, MobileNetV2, EfficientNet-B0, and a standard Vision Transformer) are trained using image data only, following their respective standard configurations, while the proposed framework is trained using the full multimodal input.

6. Evaluation Metrics

The performance of the proposed multimodal AI model is evaluated using standard classification metrics.

Accuracy

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Precision

Measures correct positive disease predictions:

$$Precision = TP / (TP + FP) \quad (2)$$

Recall

Measures disease detection capability:

$$Recall = TP / (TP + FN) \tag{3}$$

F1-score

Balances precision and recall:

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \tag{4}$$

Additional deployment metrics:

- Number of parameters
- Inference latency per image (ms)
- Model size (MB)

7. Results and Discussion

7.1 Performance Comparison

The proposed multimodal explainable lightweight transformer model is compared with conventional deep learning approaches, each trained on the image modality only, following their standard published configurations.

Table 1. Accuracy and parameter count of baseline models versus the proposed framework (single-run, image-only baselines).

Model	Accuracy (%)	Parameters
VGG16	93.4	138 M
ResNet50	95.2	25 M
MobileNetV2	96.1	3.5 M
EfficientNet-B0	96.8	5.3 M
Vision Transformer (ViT-Base)	97.2	86 M

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

Model	Accuracy (%)	Parameters
Proposed Multimodal Lightweight Transformer	98.5	6.2 M

As shown in Table 1, the proposed framework achieves improved accuracy (98.5%) while maintaining substantially lower computational complexity (6.2 M parameters) than the closest-performing baselines, including the standard Vision Transformer (97.2% accuracy, 86 M parameters). Because the baselines are unimodal, this comparison conflates architectural gains with the added information from the sensor and weather modalities; Section 7.2 isolates the fusion contribution using an image-only variant of the proposed architecture as a controlled reference point. Results reflect a single training run without multi-seed variance estimates (see Section 9).

7.2 Impact of Multimodal Data Fusion

To analyze the contribution of each modality in isolation from architectural differences, the proposed model architecture itself (not external baselines) is retrained using different input combinations, holding the network design fixed.

Table 2. Accuracy of the proposed architecture under different input-modality combinations (single-run).

Input Data	Accuracy (%)
Leaf images only	94.6
Leaf images + environmental sensor data	96.9
Leaf images + weather data	96.3
Leaf images + environmental sensor + weather data (full multimodal)	98.5

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

The ablation results show that adding environmental sensor data to leaf images yields the largest single-modality improvement (+2.3 percentage points), followed by weather data (+1.7 percentage points), with the full multimodal configuration achieving the highest accuracy of 98.5%. Because the sensor and weather values used here are synthetically generated (Section 4.2–4.3), these gains demonstrate that the fusion mechanism can exploit auxiliary signal when present, rather than confirming that real-world environmental data will yield a comparable improvement; that claim requires validation on field-collected data.

7.3 Explainability Analysis

Grad-CAM visualizations generated by the proposed model consistently localize attention on lesion regions, discoloration patterns, and leaf margins associated with the predicted disease class, rather than background or irrelevant leaf area. SHAP-based attribution over the non-visual modalities indicates that humidity and soil moisture are frequently among the most influential environmental features, consistent with known agronomic drivers of fungal and bacterial disease prevalence.

7.4 Computational Efficiency and Deployment Considerations

With 6.2 million parameters and a compact model footprint, the proposed framework is suitable for deployment on mobile and low-power edge devices, supporting real-time, in-field disease screening without reliance on continuous cloud connectivity.

8. Reproducibility and Data/Code Availability

All experiments in this preliminary study use a single fixed random seed; multi-seed replication has not yet been performed (see Section 9). Hardware specification, exact hyperparameter search procedure, and training wall-clock time are not yet finalized for reporting and will be included upon completion of full-scale experiments. Code and the synthetic sensor/weather generation procedure are intended to be released in a public repository upon acceptance; the PlantVillage image data is publicly available under its original license. Any future release using real sensor/weather data will include a data statement covering collection sites, dates, instrumentation, and licensing.

9. Limitations

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

- Synthetic auxiliary modalities: environmental sensor and weather values are synthetically generated rather than collected in the field (Sections 4.2–4.3), which limits the real-world validity of the reported multimodal fusion gains until replicated on genuine paired data.
- Single-run results: accuracy figures in Tables 1 and 2 reflect one training run each; no confidence intervals, standard deviations, or statistical significance tests (e.g., paired t-test, McNemar's test) are yet reported across multiple seeds.
- Confounded baseline comparison: Table 1 compares the proposed multimodal model against unimodal baselines, so the reported gain reflects both architectural and data-availability differences; Table 2 partially addresses this via a same-architecture ablation, but a full multimodal-vs-unimodal comparison using identical base architectures across all baselines has not yet been performed.
- No inference latency or model-size measurements: Section 6 lists these as evaluation metrics, but empirical values are not yet reported and are needed to substantiate real-time edge-deployment claims.
- Explainability results are qualitative: Grad-CAM and SHAP outputs are described narratively; no quantitative localization or attribution metrics, sample counts, or example figures are yet included.
- Dataset scope: evaluation uses PlantVillage, a curated dataset of leaf images captured under relatively controlled conditions; generalization to field-captured images with variable lighting, background clutter, and co-occurring stressors is untested.
- Missing-modality robustness at inference time is untested — a practical requirement given that sensor connectivity may be intermittent in real field deployments.

10. Conclusion and Future Work

This paper presented a multimodal, explainable, and lightweight transformer-based framework for real-time plant disease identification that integrates leaf images, environmental sensor data, and weather information. In a preliminary evaluation using synthetic auxiliary modalities, the proposed

Journal of Computer Science and Intelligent Systems
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

framework outperforms established image-only CNN and transformer baselines in classification accuracy while using substantially fewer parameters, and multimodal fusion provides measurable gains within the same architecture. The incorporation of Grad-CAM and SHAP-based explainability is intended to support transparency in model predictions, though this has so far been assessed only qualitatively.

The priority for future work is replacing the synthetic environmental and weather data with real, field-collected, geotagged sensor and weather records paired to leaf imagery, followed by: multi-seed statistical validation with significance testing; a controlled comparison against same-architecture unimodal variants and against the closely related multimodal methods identified in Section 2.5; quantitative explainability evaluation; measured on-device inference latency and model size across representative mobile hardware; and an assessment of robustness to missing or noisy sensor modalities at inference time. Longer-term extensions include broader crop and disease coverage and federated learning strategies for privacy-preserving model updates across distributed farms.

References

Ahmad, A., Saraswat, D., & El Gamal, A. (2023). A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. Smart Agricultural Technology, 3, 100083. https://doi.org/10.1016/j.atech.2022.100083

Amara, J., König-Ries, B., & Samuel, S. (2023). Concept explainability for plant diseases classification. In Proceedings of the 18th International Conference on Computer Vision Theory and Applications (VISAPP 2023) (Vol. 4, pp. 246–253). SCITEPRESS. https://doi.org/10.5220/0011667900003417

Bao, W., Yang, X., Liang, D., Hu, G., & Yang, X. (2021). Lightweight convolutional neural network model for field wheat ear disease identification. Computers and Electronics in Agriculture, 189, 106367. https://doi.org/10.1016/j.compag.2021.106367

Journal of Computer Science and Intelligent Systems
JOURNAL India Publication
Volume 1, Issue 1 | July - September 2026 | www.journalindiapublication.com | 2026 Journal India Publication

Borhani, Y., Khoramdel, J., & Najafi, E. (2022). A deep learning based approach for automated plant disease classification using vision transformer. Scientific Reports, 12, 11554. https://doi.org/10.1038/s41598-022-15163-0

De Silva, M., & Brown, D. (2023). Multispectral plant disease detection with vision transformer–convolutional neural network hybrid approaches. Sensors, 23(20), 8531. https://doi.org/10.3390/s23208531

Dhaka, V. S., Kundu, N., Rani, G., Zumpano, E., & Vocaturo, E. (2023). Role of Internet of Things and deep learning techniques in plant disease detection and classification: A focused review. Sensors, 23(18), 7877. https://doi.org/10.3390/s23187877

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318. https://doi.org/10.1016/j.compag.2018.01.009

Gole, P., Bedi, P., Marwaha, S. S., Haque, M. A., & Deb, C. K. (2023). TrIncNet: A lightweight vision transformer network for identification of plant diseases. Frontiers in Plant Science, 14, 1221557. https://doi.org/10.3389/fpls.2023.1221557

Lee, S., & Yun, C. M. (2023). A deep learning model for predicting risks of crop pests and diseases from sequential environmental data. Plant Methods, 19, 145. https://doi.org/10.1186/s13007-023-01122-x

Li, G., Wang, Y., Zhao, Q., Yuan, P., & Chang, B. (2023). PMVT: A lightweight vision transformer for plant disease identification on mobile devices. Frontiers in Plant Science, 14, 1256773. https://doi.org/10.3389/fpls.2023.1256773

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. https://doi.org/10.3389/fpls.2016.01419

Shoaib, M., Shah, B., El-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., & Ali, F. (2023). An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in Plant Science, 14, 1158933. https://doi.org/10.3389/fpls.2023.1158933

Thakur, P. S., Khanna, P., Sheorey, T., & Ojha, A. (2022). Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT. [Conference/preprint version].

Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. (2019). A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 161, 272–279. https://doi.org/10.1016/j.compag.2018.03.032