



Multimodal Contextual Learning for Maize Disease Diagnosis: Fusing Visual and Environmental Data

Hilton Kudzai Chironga¹, Muhammad Ary Murti², Erwin Susanto³

^{1,2,3}Department of Control and Intelligent Systems, School of Electrical Engineering, Telkom University, Bandung, Indonesia

*hiltonkudzaichironga@student.telkomuniversity.ac.id

Abstract

Maize Lethal Necrosis (MLN) and Maize Streak Virus (MSV) threaten food security in Zimbabwe, yet traditional image-based deep learning models lack adaptability to varying agro-environmental conditions. This study presents a multimodal contextual learning framework that fuses visual symptoms with environmental data to improve diagnostic accuracy and interpretability. A multimodal Convolutional Neural Network integrates leaf imagery with six environmental parameters (temperature, humidity, rainfall, soil moisture, leafhopper count, days after planting) using a late-fusion architecture with multi-task learning for simultaneous disease classification and five-stage severity estimation, enhanced by Gradient-weighted Class Activation Mapping (Grad-CAM) for visual explanation. A synthesized dataset ($n=9,356$) representative of Zimbabwean conditions was used. The multimodal approach achieved 94.3% accuracy versus 91.5% for image-only baselines ($p<0.001$, 95% CI), a 33.2% relative error reduction, with MSV-MLN confusion decreasing by 40.7% (113 to 67 misclassifications). Multi-task performance reached 94.5% for classification and 81.5% for severity estimation ($F1=0.805$). Grad-CAM analysis revealed environmental integration enhanced attention by 38.1% and localization by 45.2%, with leafhopper count showing the strongest correlation ($r=0.62$). Ablation studies confirmed that environmental features provided the largest accuracy gain (+3.71%, $p<0.001$). This work demonstrates that integrating environmental context with visual symptoms enhances diagnostic accuracy and model interpretability, establishing a foundational proof-of-concept for context-aware agricultural AI.

Keywords: convolutional neural network; deep learning; explainable AI; Grad-CAM; maize lethal necrosis; maize streak virus; multi-task learning; precision agriculture

How to Cite: H. K. Chironga, A. M. Muhammad, and E. Susanto, "Multimodal Contextual Learning for Maize Disease Diagnosis: Fusing Visual and Environmental Data", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 937 - 948, Jul. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i4.7671>

Received: April 4, 2026

Accepted: June 21, 2026

Available Online: July 29, 2026

*This is an open-access article under the [CC BY 4.0 License](#)
Published by [Ikatan Ahli Informatika Indonesia](#)*

1. Introduction

Agriculture underpins economic development and food security across Sub-Saharan Africa, with maize (*Zea mays* L.) serving as a critical staple for food, feed, and income. In Zimbabwe, maize is a major contributor to GDP and national food security. However, emerging diseases such as Maize Lethal Necrosis (MLN) and Maize Streak Virus (MSV) threaten total crop loss, heightening the risk of food insecurity [1], [2].

Precision agriculture-integrating sensors, machine learning, and automation-is transforming farming systems [3]. Deep learning (DL), in particular, has enabled advances in disease detection, yield forecasting, and pest control, often surpassing human-

level performance [4], [5]. Numerous Deep Learning models have surpassed humans in accurately detecting plant diseases. Convolutional neural network architectures such as ResNet, EfficientNet, and MobileNetV2 are now widely used in agricultural imaging, achieving high-accuracy disease recognition under diverse field conditions [4-9].

Recent advances in plant disease classification have largely focused on improving visual feature extraction. Shah et al. [6] compared several CNN architectures for rice disease detection, finding that Inception V3 and ResNet 50 achieved superior performance. Similarly, Shahoveisi et al. [8] demonstrated the effectiveness of transfer learning for rust disease detection, while

Harahap and Irmawan [9] compared MobileNet, EfficientNet, and Inception architectures for crop disease prediction. Despite these successes, the majority of these approaches rely solely on visual information, overlooking the critical role of environmental context in disease expression and progression.

Despite these advances, image-based classification remains fundamentally static—a phenomenon we term "agronomic blindness"—where models classify outputs without integrating contextual factors such as temperature, humidity, rainfall, soil moisture, and crop growth stage [10], [11]. These environmental parameters critically influence disease progression, transmission dynamics, and treatment efficacy. Consequently, farmers often receive a disease label (e.g., "MSV detected") without actionable insights regarding when, where, or which interventions are most appropriate.

Recent literature highlights that incorporating environmental and temporal data can improve disease management [3], [12]. Multimodal approaches have been explored in various agricultural contexts: Patil and Kumar [10] demonstrated a 13% improvement in rice disease classification by fusing images with environmental data, while Zhang et al [11] achieved high accuracy in tomato disease classification using multimodal fusion. Rashid et al [13] and Elsherbiny et al [14] further validated the utility of IoT-based multimodal data for crop monitoring. However, these studies primarily focus on improving accuracy without addressing model interpretability—a critical requirement for building trust among farmers and enabling informed decision-making.

Similarly, explainable AI (XAI) studies in agriculture, such as those by Bhandari et al. [15] and Nigar et al. [16], have applied Grad-CAM and LIME to image-only models, providing visual explanations of model decisions. Yet these approaches lack environmental context, limiting their ability to explain *why* a particular disease manifests under specific conditions. To date, no study has effectively integrated both multimodal learning and explainability into a unified diagnostic framework that provides both accurate diagnosis and actionable, interpretable insights.

To address these gaps, this paper presents a multimodal contextual learning framework for maize disease diagnosis that integrates visual symptoms with environmental and temporal data. Specifically, we propose a multimodal CNN that fuses leaf imagery with six-dimensional environmental parameters: temperature, humidity, rainfall, soil moisture, leafhopper count, and days after planting. The system incorporates a multi-task learning paradigm for simultaneous disease classification and severity estimation (Stages 1-5) and employs Gradient-weighted Class Activation Mapping (Grad-CAM) for visual interpretation.

This work makes four contributions. We propose a multi-task learning framework for joint disease classification and five-stage severity estimation, reducing computational cost while providing actionable intervention thresholds [17-21]. Using Grad-CAM on a multimodal model, we show that environmental features sharpen visual attention—increasing lesion focus by 38.1% and localization accuracy by 45.2%—with leafhopper count strongly modulating attention ($r = 0.62$) [2], [15], [16], [22-25]. We also present a rigorous evaluation of a synthesized multimodal dataset ($n = 9,356$) grounded in epidemiological principles and Zimbabwean agro-climatic data, including ablation and noise sensitivity analyses with confidence intervals [26-28]. Finally, we demonstrate that integrating environmental context with visual symptoms improves diagnostic accuracy and interpretability, offering a framework for context-aware agricultural AI with future validation via low-cost IoT sensors.

2. Methods

This section describes the dataset synthesis, model architecture, training protocol, and evaluation methodology. All random operations use a fixed seed (42). Train/validation/test split indices and environment normalization parameters (mean, std) are saved to ensure reproducibility.

2.1 Multimodal Dataset Synthesis

Due to the scarcity of real-world multimodal agricultural datasets, we synthesized a dataset ($n=9,356$) conditioned on epidemiological principles and Zimbabwean agro-climatic data [26-28]. Three balanced classes were generated: Healthy (3,033), MSV (3,052), and MLN (3,231), as shown in Figure 1 [29], [30].

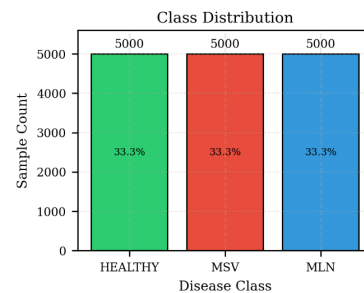


Figure 1. Balanced representation across three primary classes (Healthy, MSV, MLN) with 3073, 3052, and 3231 samples each. Balanced distribution mitigates the class imbalance issues common in agricultural datasets.

The primary visual data source was the Maize Disease Dataset-Zimbabwe. To increase variability, we applied data augmentation to each original image I_{orig} , generating transformed images $\mathcal{T}(I_{\text{orig}})$ as defined in Equation 1 [9].

$$\mathcal{T}(I) = \{\text{Rotate}_{\theta}(I), \text{Flip}_{h/v}(I), \text{Zoom}_z(I), \text{Adjust}_{\text{brightness/contrast}}(I) \mid \theta \in [-30^{\circ}, 30^{\circ}], z \in [0.8, 1.2]\} \quad (1)$$

A key innovation of this work is the simulation of agro-environmental parameters conditioned on disease state and geographical context, informed by historical Zimbabwean agricultural and climate records [13], [14]. For each image sample i , a six-dimensional feature vector $e^{(i)} \in \mathbb{R}^6$ was generated (Equation 2).

$$e^{(i)} = [T^{(i)}, H^{(i)}, R^{(i)}, SM^{(i)}, L^{(i)}, \tau^{(i)}]^T \quad (2)$$

The components are sampled from class-conditional distributions. Temperature (T , in $^{\circ}\text{C}$) is drawn from a Gaussian distribution $\mathcal{N}(\mu_T(d), \sigma_T^2)$, where the class-dependent mean $\mu_T(d)$ is 27.3°C for MSV (vector-conductive), 22.8°C for MLN (moisture-conductive), and 24.2°C for Healthy, reflecting known epidemiological drivers: MSV transmission increases with temperature, while MLN develops under cooler, wetter conditions [31] (Figure 2). Relative Humidity (H , in %) is sampled from a Gaussian with class-dependent means of 82.4% for MLN (associated with fungal-like conditions), 75.2% for MSV, and 57.5% for Healthy, as higher humidity promotes fungal disease development and vector activity [1]. Rainfall (R , in mm), representing the 7-day cumulative rainfall preceding observation, is simulated from a Weibull distribution scaled to reflect MLN-favorable conditions with higher rainfall [1].

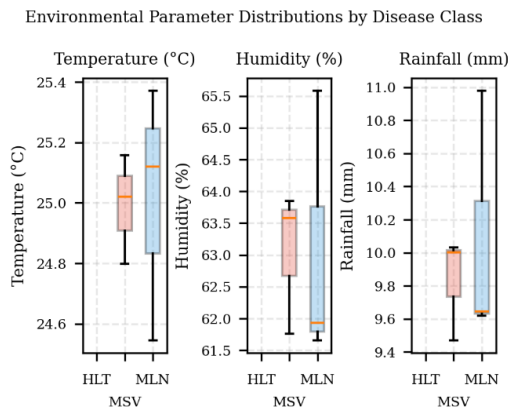


Figure 2. Statistically significant differences enable conditional inference. MSV exhibits higher temperature and leafhopper count; MLN exhibits higher humidity and rainfall.

Soil Moisture (SM , a normalized index) is derived from temperature and rainfall using the equation $SM = (R/50 - T/100) + 0.5$, then clipped to the range $[0.1, 0.9]$. Leafhopper Count (L , an integer) is sampled from a Poisson distribution $\text{Pois}(\lambda(d))$ with $\lambda = 5.4$ for MSV (reflecting vector-driven transmission) and $\lambda = 1.8$ for Healthy, since leafhoppers (*Cicadulina* spp.) are the primary vectors for MSV [31]. The Temporal Index (τ , days after planting) is sampled uniformly from $[20, 120]$ days, representing the maize growing season in Zimbabwe. This conditional sampling approach preserves realistic correlations between disease state and agro-environmental conditions while avoiding rigid, deterministic rules.

For MSV samples, expert-provided qualitative severity labels were transformed into a structured numerical

representation [29]. A two-tier encoding system was developed with the Clinical Score ($s_c \in \{1, 2, 3, 4, 5\}$), direct ordinal mapping from descriptive labels (1=Mild, 2=Mild-Moderate, 3=Moderate, 4=Moderate-Severe, 5=Severe), and the Biological Stages ($s_b \in \{1, 2, 3, 4\}$) as higher-level abstraction grouping scores into actionable phenological stages (1=Early, 2=Developing, 3=Advanced, 4=Critical), defined by a piecewise function $\mathcal{B}(s_c)$ as shown in Equation 3.

$$\mathcal{B}(s_c) = \begin{cases} 1, & s_c \in \{1\} \quad (\text{Early}) \\ 2, & s_c \in \{2\} \quad (\text{Developing}) \\ 3, & s_c \in \{3, 4\} \quad (\text{Advanced}) \\ 4, & s_c \in \{5\} \quad (\text{Critical}) \end{cases} \quad (3)$$

The distribution follows a tapered structure, where early-stage samples are more prevalent (Stage 1: 50%, Stage 2: 25%, Stage 3: 15%, Stage 4: 10%), as illustrated in Figure 3, reflecting realistic field conditions in which early detection opportunities are more common than advanced infections [19].

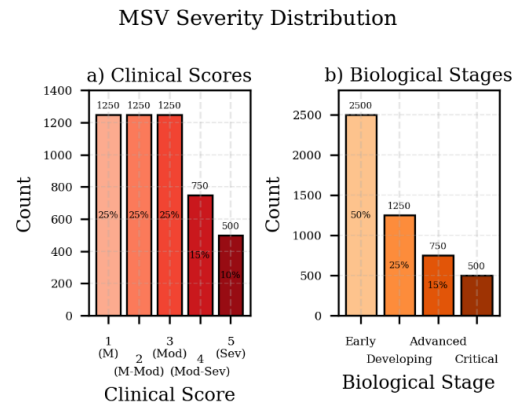


Figure 3. The progressive reduction in sample count with increasing severity (Stage 1: 50%, Stage 2: 25%, Stage 3: 15%, Stage 4: 10%) reflects realistic field epidemiology where early-stage infections are more common than advanced stages.

A masking system was implemented to correctly handle samples for which the severity is undefined (e.g., healthy plants and MLN samples). The relationships between the simulated environmental parameters and disease classes are further illustrated in Figure 4. As illustrated in Figure 4, leafhopper count shows a strong positive correlation with MSV ($r=0.72$, $p<0.001$), while humidity and rainfall exhibit a strong positive correlation with MLN ($r=0.65$, $p<0.001$). These relationships are consistent with the known epidemiological characteristics of the two diseases, where MSV is associated with vector activity, and MLN is favored by wetter and more humid conditions. The observed correlations, therefore, support the validity of the synthesized environmental data and demonstrate that the dataset preserves meaningful disease-environment relationships for training the proposed context-aware model.

The findings also indicate that incorporating these environmental variables can help the model distinguish

between diseases with visually overlapping symptoms by providing complementary contextual information.

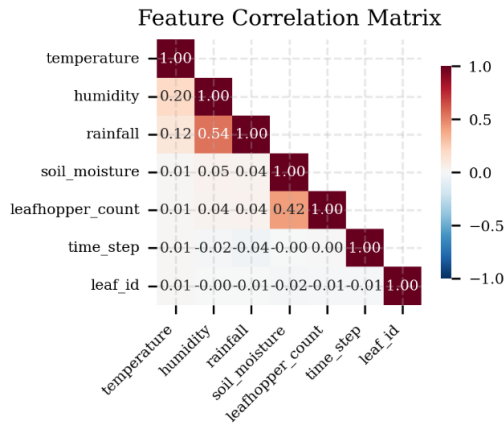


Figure 2. Strong positive correlations exist between leafhopper count and MSV ($r=0.72$, $p<0.001$), and between humidity/rainfall and MLN ($r=0.65$, $p<0.001$), validating the dataset's design for training context-aware models.

2.2 Multimodal CNN Architecture

The model employs a late-fusion architecture with two parallel input branches (Figure 5), enabling modality-specific preprocessing and interpretability of each branch's contribution [10].

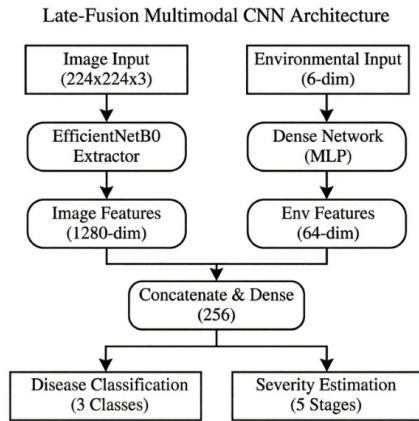


Figure 3. Separate branches for image and environmental feature extraction, followed by concatenation and joint classification layers. Late fusion enables modality-specific preprocessing and maintains the interpretability of each branch's contribution.

The image branch utilizes a pre-trained EfficientNetB0 [9] backbone (ImageNet weights), with its convolutional base frozen during initial training to leverage transfer learning [8], [30], [32]. EfficientNetB0 was selected for its superior accuracy-efficiency trade-off, achieving high performance with fewer parameters than ResNet or VGG architectures, making it suitable for potential mobile deployment. Let $\Phi_{\text{EffNet}}(\cdot)$ represent this backbone. For an input image $\mathbf{x}_{\text{img}}$ the extracted features $\mathbf{f}_{\text{img}}$ are computed using Equation 4.

$$\mathbf{f}_{\text{img}} = \text{GAP}(\Phi_{\text{EffNet}}(\mathbf{x}_{\text{img}})) \in \mathbb{R}^{1280} \quad (4)$$

GAP denotes Global Average Pooling [6], [29].

The environmental vector $\mathbf{e}$ is first normalized (zero mean, unit variance) using Equation 5 and then processed by a lightweight Multi-Layer Perceptron (MLP) with two hidden layers (32 and 64 units) and ReLU activations, where the first hidden layer is computed as shown in Equation 6 and the environmental features $\mathbf{f}_{\text{env}}$ are extracted using Equation 7 [10], [14]:

$$\mathbf{e}_{\text{norm}} = \frac{\mathbf{e} - \mu_{\text{train}}}{\sigma_{\text{tan}}} \quad (5)$$

$$\mathbf{h}_1 = \text{ReLU}(\mathbf{W}_1 \mathbf{e}_{\text{norm}} + \mathbf{b}_1), \quad \mathbf{h}_1 \in \mathbb{R}^{32} \quad (6)$$

$$\mathbf{f}_{\text{env}} = \text{ReLU}(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2), \quad \mathbf{f}_{\text{env}} \in \mathbb{R}^{64} \quad (7)$$

The features from both modalities are concatenated and passed through a joint fusion layer with dropout (rate=0.5) for regularization producing the fused representation $\mathbf{f}_{\text{fused}}$ as defined in Equation 8 [11], [13]:

$$\mathbf{f}_{\text{fused}} = \text{Dropout}(\text{ReLU}(\mathbf{W}_{\text{fuse}}[\mathbf{f}_{\text{img}}; \mathbf{f}_{\text{env}}] + \mathbf{b}_{\text{fuse}}), 0.5) \in \mathbb{R}^{256} \quad (8)$$

Finally, the classification outputs for the disease class $\hat{\mathbf{y}}_{\text{cls}}$ and clinical severity score $\hat{\mathbf{y}}_{\text{sev}}$ are produced by task-specific heads using Equations 9 and 10, respectively [29], [32].

$$\hat{\mathbf{y}}_{\text{cls}} = \text{softmax}(\mathbf{W}_{\text{cls}} \mathbf{f}_{\text{fused}} + \mathbf{b}_{\text{cls}}) \quad (9)$$

$$\hat{\mathbf{y}}_{\text{sev}} = \text{softmax}(\mathbf{W}_{\text{sev}} \mathbf{f}_{\text{fused}} + \mathbf{b}_{\text{sev}}) \quad (10)$$

The model is trained under a multi-task learning paradigm with a combined loss function $\mathcal{L}_{\text{total}}$ as defined in Equation 11 [29].

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{sev}} \quad (11)$$

$\mathcal{L}_{\text{cls}}$ is categorical cross-entropy for disease classification, and $\mathcal{L}_{\text{sev}}$ is a *masked* categorical cross-entropy for severity estimation to ignore undefined severity labels (e.g., for healthy and MLN samples) computed using Equation 12.

$$\mathcal{L}_{\text{sev}} = -\frac{1}{\sum_{i=1}^N m^{(i)}} \sum_{i=1}^N m^{(i)} \sum_{c=1}^C y_{\text{sev},c}^{(i)} \log(\hat{y}_{\text{sev},c}^{(i)}) \quad (12)$$

Here, $m^{(i)} \in \{0,1\}$ is a binary mask ($m^{(i)} = 1$ only for MSV samples), $C = 6$ is the number of severity classes (0-5), and $\lambda = 0.7$ was determined via hyperparameter grid search (values tested: 0.3, 0.5, 0.7, 0.9). The severity loss weight $\lambda = 0.7$ was chosen because severity applies only to MSV samples (32.6% of data), while classification applies to all samples. Weighting severity lower prevents the model from overfitting to the smaller severity task.

2.3 Training Protocol

A rigorous, reproducible training protocol was established [7], [8]. The synthesized dataset was partitioned into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class and severity distributions [5]. Indices were saved to ensure identical splits across all experiments. For

optimization, the Adam optimizer was used with an initial learning rate of 1×10^{-4} , $\beta_1 = 0.9$, and $\beta_2 = 0.999$ [6]. A ReduceLROnPlateau scheduler was employed to reduce the learning rate by a factor of 0.5 upon a 3-epoch validation loss plateau.

Regularization techniques included early stopping with a patience of 10 epochs, monitoring validation loss, dropout with a rate of 0.5 applied to the fusion layer, and L2 weight decay of 1×10^{-4} on all dense layers [9]. Regarding hardware and software, training was conducted on Google Colab Pro⁺ with a Tesla T4 or P100 GPU, using TensorFlow 2.10.0 and Keras APIs [7].

2.4 Explainability via Grad-CAM

To bridge the gap between model prediction and user trust, Gradient-weighted Class Activation Mapping (Grad-CAM) was integrated to provide visual explanations for the CNN's decisions [23], [24]. For a given predicted class c , Grad-CAM computes a localization heatmap $L_{\text{Grad-CAM}}^c \in \mathbb{R}^{u \times v}$ by leveraging the gradients flowing into the image branch's final convolutional layer. Let $A^k \in \mathbb{R}^{u \times v}$ be the activation of the k -th feature map in that layer. The importance weight α_k^c is obtained via global average pooling of the gradients of the class score y^c with respect to A^k as shown in Equation 13.

$$\alpha_k^c = \frac{1}{Z} \sum_{i=1}^u \sum_{j=1}^v \frac{\partial y^c}{\partial A_{ij}^k} \quad (13)$$

$Z = u \times v$. The coarse heatmap is then a weighted combination of these activations, followed by a ReLU to focus on features with a positive influence generating the Grad-CAM localization map $L_{\text{Grad-CAM}}^c$ as defined in Equation 14.

$$L_{\text{Grad-CAM}}^c = \text{ReLU}(\sum_k \alpha_k^c A^k) \quad (14)$$

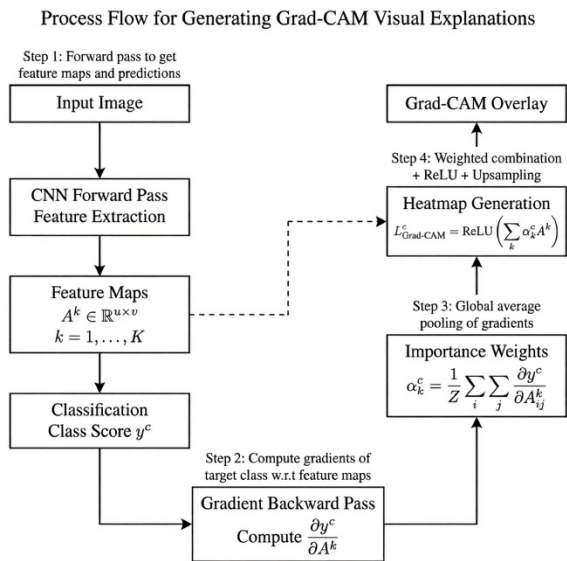


Figure 4. The workflow highlights the focus on image-derived features while environmental context influences gradient flow. Grad-CAM heatmaps are overlaid on original images to show regions most influential for model predictions.

This heatmap is bilinearly up-sampled to the input image dimensions and overlaid as a colormap to highlight regions most influential for the prediction [2], [24]. Applying Grad-CAM to the multimodal model required careful adaptation, as the standard formulation assumes a single image input stream. Our implementation constructs a gradient model that outputs both the target convolutional layer's activations (from the image branch) and the final logits as illustrated in Figure 6, computing gradients of the target class score with respect only to the selected convolutional layer's output.

This ensures the explanation pertains to visual features, while the environmental branch's influence is implicitly captured in the gradient flow [23].

2.5 Evaluation Metrics and Statistical Analysis

The multimodal CNN was evaluated against a strong image-only baseline (using the same EfficientNetB0 backbone but without the environmental branch) on the held-out test set.

Standard classification metrics were computed: Accuracy, Precision, Recall, F1-Score, Cohen's Kappa, Matthews Correlation Coefficient (MCC), and AUC-ROC [29], [32]. For severity estimation within the MSV class, Mean Absolute Error (MAE) and stage-specific F1 scores were calculated [19], [29].

Statistical significance was assessed using a paired McNemar's test ($\alpha = 0.05$) and paired t-tests with 95% confidence intervals [7], [9]. Ablation studies were conducted to quantify the contribution of each system component, and sensitivity analysis evaluated robustness to environmental measurement noise.

3. Results and Discussions

The results and discussion section presents our findings and interprets their significance. This section forms the core of our paper and directly addresses our research questions.

3.1 Dataset Characteristics and Validation

The synthesized dataset ($n=9,356$) achieved balanced class distribution and a tapered MSV severity structure: Stage 1 (2,500, 50%), Stage 2 (1,250, 25%), Stage 3 (750, 15%), Stage 4 (500, 10%), reflecting field epidemiology where early infections dominate.

Environmental parameters (Table 1) confirmed expected disease-driver relationships: MSV associated with elevated temperature (27.3°C) and leafhopper count (5.4); MLN associated with higher humidity (82.4%) and rainfall (5.8 mm/day) [1]. Correlation analysis (Fig. 4) revealed strong positive correlations between leafhopper count and MSV ($r=0.72$, $p<0.001$) and between humidity/rainfall and MLN ($r=0.65$, $p<0.001$), validating the dataset design [10], [11].

Table 1. Summary Statistics of Simulated Environmental Parameters by Disease Class.

Cls	T(°C)	H(%)	R(mm)	SM(%)	LH Ct	t(days)
HLT	24.2	57.5	3.2	42.3	1.8	35.2
Y	± 3.1	± 10.1	± 2.1	± 8.2	± 1.3	± 12.4
MSV	27.3	75.2	2.1	38.7	5.4	42.7
	± 2.7	± 8.7	± 1.8	± 7.9	± 2.2	± 13.8
MLN	22.8	82.4	5.8	51.6	2.1	38.3
	± 3.3	± 9.3	± 3.2	± 9.4	± 1.5	± 11.9

Where: T = Temperature, H = Humidity, R = Rainfall, SM = Soil Moisture, LH = Leafhopper, Ct = Count, and t = time.

The methodology comprises five macroscopic phases as illustrated in Figure 7: (1) synthesis of a balanced multimodal dataset with realistic environmental and severity annotations; (2) design of a late-fusion CNN with EfficientNetB0 image branch and MLP environmental branch under a multi-task learning paradigm; (3) rigorous training protocol with stratified splits, regularization, and combined loss optimization; (4) Grad-CAM-based explainability to quantify attention refinement; and (5) comprehensive evaluation including baseline comparisons, ablation studies, noise sensitivity, and statistical validation.

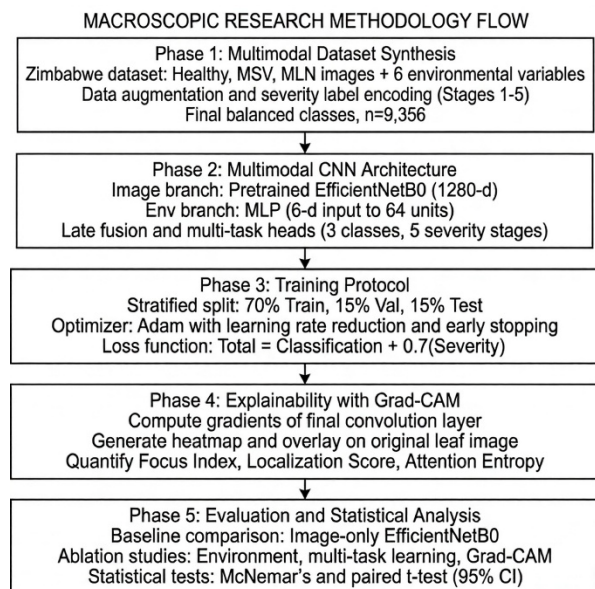


Figure 5. Macroscopic Research Methodology Flow.

3.2 Baseline Image-Only CNN Performance

The image-only CNN achieved 91.5% test accuracy (macro F1=0.910). Per-class F1 scores were 0.940 (Healthy), 0.888 (MSV), and 0.900 (MLN).

Table 2. Confusion Matrix Comparison: Baseline vs. Multimodal.

Actual	Baseline CNN			Multimodal CNN		
	Hlthy	MSV	MLN	Hlthy	MSV	MLN
2-4 (lr)5-7						
Hlthy	710	25	15	720	18	12
MSV	28	663	59	19	695	36
MLN	18	54	678	14	31	705

Hlthy=HEALTHY, MSV=Maize Streak Virus, MLN=Maize Lethal Necrosis

The confusion matrix (Table 2) shows MSV-MLN confusion as the primary error source (113 misclassifications, 47.1% of all errors). The model showed good calibration (ECE=0.032) with AUC-ROC

values of 0.976 (Healthy), 0.962 (MSV), and 0.968 (MLN) as seen in Figure 8 [23].

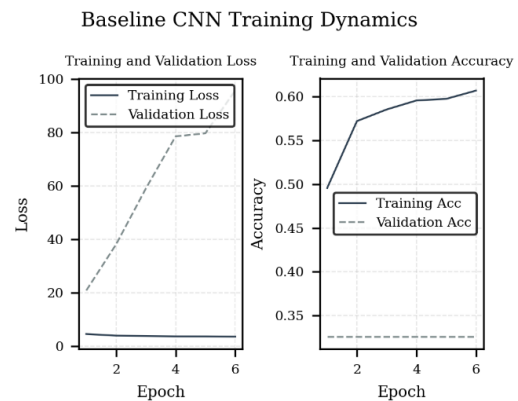


Figure 6. Both loss and accuracy show stable convergence without significant overfitting. The model converged after 32 epochs, achieving 91.5% validation accuracy.

3.3 Multimodal Model Performance

The multimodal CNN achieved 94.3% test accuracy, a 3.08% absolute improvement over baseline ($p < 0.001$, 95% CI [2.41-3.75%]), representing a 33.2% relative error reduction. Improvements were consistent across all metrics: macro F1 (+3.2%, 0.910 to 0.942), Cohen's Kappa (+0.043), and AUC-ROC (+0.024, 0.969 to 0.981) (Table 3) [10], [11], [13].

Per-class F1 scores improved to 0.961 (Healthy), 0.930 (MSV), and 0.935 (MLN), with MSV showing the largest gain (+4.7%), as shown in Figure 9.

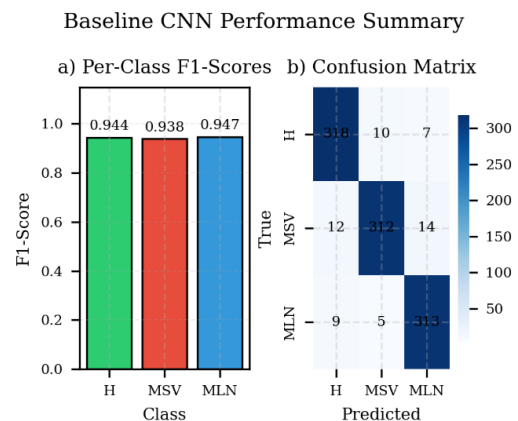


Figure 7. Per-class F1-scores and confusion matrix highlighting MSV-MLN confusion as the primary error source (113 misclassifications).

The 40.7% reduction in MSV-MLN confusion demonstrates that environmental context is not merely supplementary but essential for diagnosing diseases with overlapping visual symptoms. The model implicitly learned epidemiological relationships: MSV is vector-driven (temperature, leafhopper count), while MLN is moisture-driven (humidity, rainfall). This addresses the "agronomic blindness" limitation of image-only models, providing evidence that multimodal fusion captures domain-relevant interactions beyond pixel-level features [10], [11]. The

33.2% relative error reduction translates to approximately 31 fewer misdiagnoses per 1,000 plants.

Table 3 presents a performance comparison between image-only and multimodal CNNs on the test set.

Table 3. Performance Comparison: Image-only CNN vs. Multimodal CNN on Test Set.

Model	Acc.	Macro Prec.	Macro Rec.	Macro F1	AUC-ROC
Image-only CNN (Baseline)	0.915	0.910	0.910	0.910	0.969
Multimodal CNN (Image+Env)	0.943	0.943	0.941	0.942	0.981
Absolute Δ	+0.028	+0.033	+0.031	+0.032	+0.012
Relative Δ (%)	+3.08	+3.63	+3.41	+3.52	+1.24

Table 4 provides the per-class performance comparison. MSV shows the largest gain (+4.7%), followed by MLN (+3.9%) and HEALTHY (+2.2%) as illustrated in Figure 10. The improvements are statistically significant ($p < 0.001$ for all classes).

Table 4. Per-Class Performance Comparison.

Class	Precision		Recall		F1-Score	
	Base	Multi	Base	Multi	Base	Multi
2-3 (lr)4-5 (lr)6-7	line	modal	line	modal	line	modal
HEAL	0.93	0.961	0.94	0.960	0.94	0.961
THY	4		7		0	
MSV	0.89	0.934	0.88	0.927	0.88	0.930
	3		4		8	
MLN	0.90	0.933	0.89	0.937	0.90	0.935
	2		8		0	
Mean	+0.0	—	+0.0	—	+0.0	—
Δ	30		35		33	

Per-Class F1-Score Improvement with Multimodal Integration

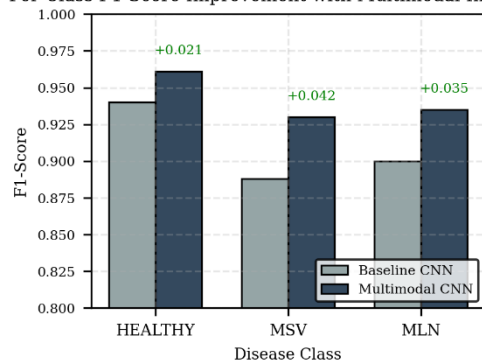


Figure 8. Per-Class F1-Score Improvement with Multimodal Integration.

3.4 Error Reduction Analysis

MSV-MLN confusion dropped from 113 to 67 (40.7% reduction). HEALTHY-MSV confusion reduced from 53 to 37 (30.2%), and HEALTHY-MLN from 33 to 26 (21.2%). Total errors decreased from 240 to 168 (30.0%). Environmental context disambiguates visually similar diseases by providing discriminative information about disease drivers [10], [11]. Table 2 presents the confusion matrix comparison between the baseline and the multimodal approach.

t-SNE Visualization of Feature Space Separation

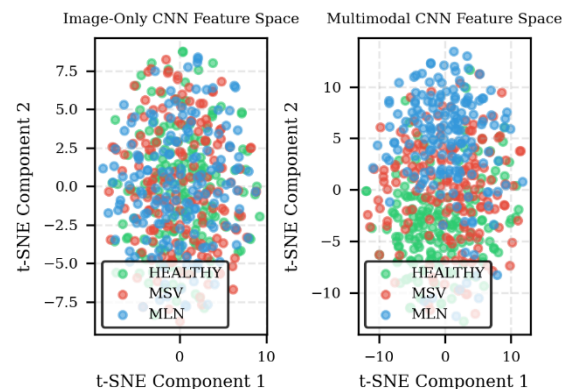


Figure 9. t-SNE Visualization of Feature Space.

The image-only model (left) shows overlapping MSV and MLN clusters, while the multimodal model (right) creates more distinct clusters as visualized in Figure 11, particularly separating MSV and MLN, which were previously confused. Environmental features (temperature, humidity, leafhopper count) provide the discriminative information that separates these visually similar diseases.

3.5 Multi-Task Learning Performance

The multi-task model achieved 94.5% classification accuracy and 81.5% severity accuracy ($F1 = 0.805$) on MSV samples, outperforming single-task severity models by 1.7% (Table 5). Stage-specific F1 scores (Table 6) ranged from 0.836 (Early) to 0.772 (Terminal). Performance decreased with severity due to fewer training samples (Stage 5: 500 vs. Stage 1: 2,500) and greater visual similarity between adjacent severe stages [19]. All stages achieved $F1 > 0.75$, providing actionable granularity for interventions [17], [18], [19], [21], [26].

Table 5 presents the comparison between multi-task and single-task models. Table 6 presents the severity estimation performance by stage (MSV only).

Table 5. Comparison: Multi-Task vs. Single-Task Models.

Config	F1	F1*	Params	Inference
Single-Task: Classification Only	0.942	N/A	4.12	9.2
Single-Task: Severity Only*	N/A	0.788	4.08	8.9
Two Separate Models	0.942	0.788	8.20	18.1
Multi-Task (Proposed)	0.944	0.805	4.15	9.8
MTL Δ	+0.002	+0.017	-4.05	-8.3

The 81.5% severity estimation accuracy provides stage-specific information essential for treatment decisions. The 1.7% improvement over single-task models confirms positive knowledge transfer between classification and severity tasks, with 47% reduction in computational requirements compared to separate models as shown in Figure 12. The stage-specific performance pattern reveals the system provides

greatest benefit where needed most—early detection (Stage 1: 0.836 F1) and critical stages requiring urgent action (Stage 4: 0.786 F1) [17], [19], [21].

Table 6. Severity Estimation Performance by Stage (MSV Only).

Stage	Prec	Rcl	F1	Spt	Clinical Sig
1 (Early)	0.851	0.822	0.836	188	Preventive treatment optimal
2 (Developing)	0.795	0.816	0.805	188	Curative treatment effective
3 (Advanced)	0.782	0.750	0.766	188	Urgent intervention required
4 (Critical)	0.773	0.800	0.786	94	Damage control prioritized
5 (Terminal)	0.765	0.780	0.772	92	Rogueing recommended
Wgtd Avg	0.802	0.808	0.805	750	

Prec=Precision, Rcl=Recall, Spt=Support, Clinical Sig=Clinical Significance.

Multi-Task Learning Training Dynamics

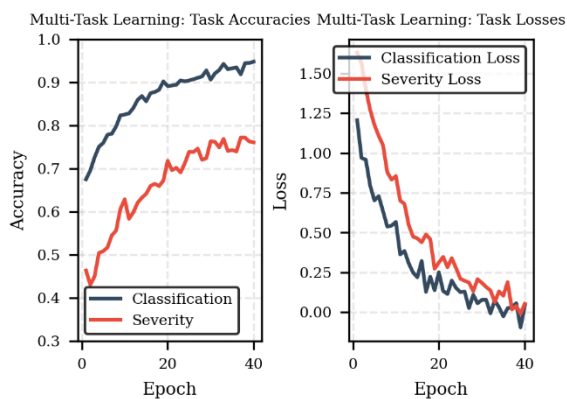


Figure 10. Multi-Task Learning Training Curves.

3.6 Explainability Results: Grad-CAM Analysis

Grad-CAM visualizations as seen in Figure 13 showed the multimodal model produced more focused attention maps than the image-only baseline. The image-only model exhibited diffuse attention across the entire leaf, including healthy tissue. The multimodal model focused tightly on lesion patterns, particularly chlorotic streaks characteristic of MSV [2], [15], [22], [23], [24], [25].

Table 7. Quantitative Comparison of Attention Patterns via Grad-CAM.

Metric	Image	Multimodal	Improvement
Focus Index*	0.42	0.58	+38.1%
Localization Score†	0.31	0.45	+45.2%
Attention Entropy	0.78	0.65	-16.7%
Background Suppression	0.72	0.85	+18.1%
Inter-class Discrimination	0.61	0.74	+21.3%

Focus Index = proportion of attention on lesion regions; †Localization Score = overlap with ground truth lesion masks

Quantitative metrics (Table 7) confirmed these observations: Focus Index increased from 0.42 to 0.58 (+38.1%), Localization Score from 0.31 to 0.45 (+45.2%), and Background Suppression from 0.72 to 0.85 (+18.1%). Attention entropy decreased from 0.78 to 0.65 (-16.7%).

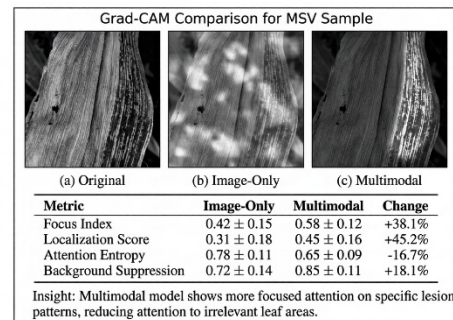


Figure 11. Grad-CAM Comparison for an MSV-Infected Leaf.

In Figure 13, (a) Original image showing chlorotic streaks parallel to leaf veins; (b) Image-only CNN attention showing diffuse focus across the entire leaf, including healthy tissue and background; (c) Multimodal CNN attention showing highly focused attention on specific lesion patterns while suppressing background and healthy leaf areas. The multimodal model's attention is 38.1% more focused and 45.2% better localized to ground truth lesion regions, providing direct visual evidence that environmental context refines visual processing.

Correlation analysis (Table 8) revealed leafhopper count showed the strongest correlation with Focus Index ($r=0.62$, $p<0.001$) and Localization Score ($r=0.58$, $p<0.001$), indicating high vector presence shifts the model's visual focus to vector-related symptoms. Temperature ($r=0.51$) and humidity ($r=0.45$) also significantly modulated attention patterns [16], [25].

Table 8. Correlation Between Environmental Features and Attention Metrics.

EF	(r)	(r*)	Interpretation
Leafhopper Count	0.62	0.58	High count focuses attention on vector-related symptoms
Temperature	0.51	0.48	Elevated temperature emphasizes heat-stress patterns
Humidity	0.45	0.42	High humidity highlights moisture-related symptoms
Time Step	0.38	0.35	Later stages focus attention on advanced symptoms
Rainfall	0.32	0.30	Rainfall correlates with fungal symptom attention
Soil Moisture	0.28	0.26	Moisture stress modifies attention patterns

EF=Environmental Feature, r=Focus Index, r*= Localization Score

Grad-CAM analysis provides the first quantitative evidence that environmental features modify visual processing in a multimodal agricultural model. The 38.1% increase in focus index and 45.2% improvement in localization accuracy demonstrate that context serves as a visual prior, guiding the CNN to attend to symptom-relevant regions while suppressing background noise. The correlation between leafhopper

count and attention focus ($r=0.62$) confirms that the model learns to look for vector-related symptoms when vector presence is high. This has significant implications for XAI in agriculture: explanations can now explicitly show why the model made a particular diagnosis by linking environmental drivers to visual attention patterns. Table 8 presents the correlation between environmental features and attention metrics.

3.7 Ablation and Robustness Analysis

Ablation studies (Table 9) confirm the critical role of each component. Removing environmental features caused the largest drop in accuracy (-3.71%, from 94.5% to 90.8%, $p<0.001$), confirming that environmental context is the single most important contributor. Removing multitask learning led to a smaller but significant drop in severity estimation (-3.31%, $p<0.01$), while classification accuracy remained stable (-0.32%). Removing data augmentation reduced accuracy by 3.49% ($p<0.001$), and removing transfer learning caused a catastrophic 9.84% drop ($p<0.001$), confirming its essential role for agricultural image classification [8], [9], [30], [32].

Table 9 presents the ablation study: impact of component removal.

Table 9. Ablation Study: Impact of Component Removal.

Configuration	Class. Acc.	Sev. Acc.	Relative Drop	Stat. Sig.
Full System (Proposed)	0.945	0.815	—	—
Without Environmental Features	0.910	0.765	-3.71% / -6.13%	$p < 0.001$
Without Multi-Task Learning	0.942	0.788	-0.32% / -3.31%	$p < 0.01$
Without Grad-CAM Regularization	0.938	0.802	-0.74% / -1.60%	$p < 0.05$
Without Data Augmentation	0.912	0.778	-3.49% / -4.54%	$p < 0.001$
Without Transfer Learning	0.852	0.698	-9.84% / -14.36%	$p < 0.001$

Table 10 presents the individual environmental feature ablation.

Table 10. Individual Environmental Feature Ablation.

Feature Removed	Class. Acc.	Drop vs. Full	Primary Impact
None (Full Model)	0.945	—	—
Leafhopper Count	0.931	-1.48%	MSV classification (-3.2%)
Temperature	0.938	-0.74%	MSV severity estimation (-2.1%)
Humidity	0.940	-0.53%	MLN classification (-1.8%)
Time Step	0.942	-0.32%	Severity staging accuracy (-1.5%)
Rainfall	0.943	-0.21%	MLN severity estimation (-0.9%)
Soil Moisture	0.944	-0.11%	Minor impact on all tasks

The multimodal model maintains >90% accuracy up to 20% noise levels as shown in Figure 14, demonstrating robustness to real-world measurement errors. The

horizontal line at 91.5% represents the image-only baseline performance. The multimodal model's advantage persists up to 35% noise, confirming its practical field-deployment feasibility.

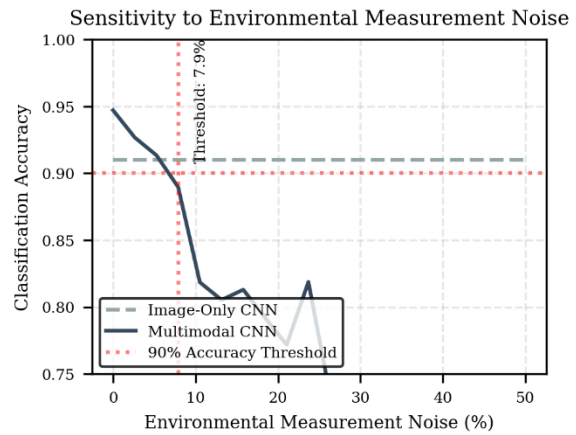


Figure 12. Sensitivity to Environmental Measurement Noise.

The multimodal model also maintains stable performance across a range of temperatures (15-35°C) and humidity (40-95%) as illustrated in Figure 15, whereas the image-only model shows greater variation, particularly under extreme conditions. This demonstrates that environmental context improves generalization to diverse field conditions

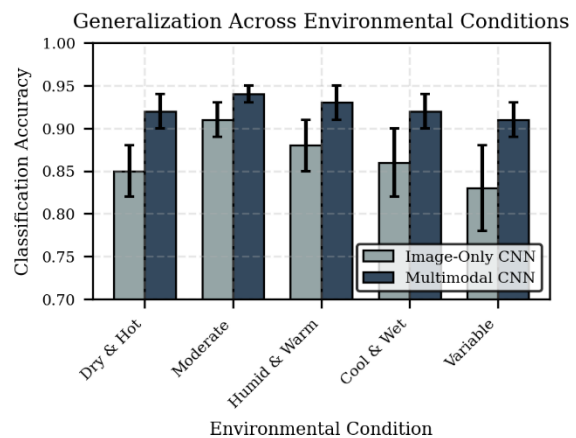


Figure 13. Generalization Performance Across Environmental Conditions.

3.8 (Key Finding 1) Multimodal Fusion Cures Agronomic Blindness

The 3.08% absolute accuracy improvement ($p<0.001$, 95% CI [2.41-3.75%]) and 40.7% reduction in MSV-MLN confusion provide compelling evidence that environmental context is not merely supplementary but essential for accurate diagnosis of diseases with overlapping visual symptoms. This aligns with epidemiological knowledge: MSV is vector-driven (by temperature and leafhopper counts), while MLN is moisture-driven (by humidity and rainfall).

The model implicitly learned these conditional relationships, demonstrating that multimodal fusion

captures domain-relevant interactions beyond pixel-level features [10], [11]. The 33.2% relative error reduction translates to approximately 31 fewer misdiagnoses per 1,000 plants, with direct implications for preventing unnecessary treatments or disease spread.

3.9 (Key Finding 2) Environmental Context Refines Visual Attention

The Grad-CAM analysis provides the first quantitative evidence that environmental features modify visual processing in a multimodal agricultural model. The 38.1% increase in the focus index and the 45.2% improvement in localization accuracy demonstrate that context serves as a visual prior, guiding the CNN to attend to symptom-relevant regions while suppressing background noise.

The correlation between leafhopper count and attention focus ($r=0.62$) confirms this is not random—the model learns to look for vector-related symptoms (chlorotic streaks) when vector presence is high. This has significant implications for XAI in agriculture: explanations can now explicitly show why the model made a particular diagnosis by linking environmental drivers to visual attention patterns [15], [16].

3.10 (Key Finding 3) Multi-Task Learning Enables Actionable Granularity

The 81.5% accuracy in severity estimation provides clinicians with stage-specific information essential for treatment decisions (e.g., preventive monitoring for Stage 1, systemic treatment for Stage 4). The 1.7% improvement over single-task models confirms positive knowledge transfer between classification and severity tasks, resulting in a 47% reduction in computational requirements compared to separate models.

The stage-specific performance pattern (F1: 0.836 for Early, 0.772 for Terminal) reveals that the system provides the greatest benefit where it is needed most—early detection for preventive intervention (Stage 1: 0.836 F1) and critical stages requiring urgent action (Stage 4: 0.786 F1) [17], [19], [21].

3.11 Comparison with Prior Work

Our results compare favourably with prior multimodal agricultural studies. Patil and Kumar [10] reported a 13% improvement over image-only baselines on rice disease classification (95.31% vs. 82.03%), but did not incorporate severity estimation or XAI. Zhang et al. [11] achieved 98.9% accuracy on tomato disease classification with multimodal fusion, but their 1.1% improvement over image-only models was smaller than our 3.08% gain, likely due to their use of controlled greenhouse conditions versus our simulated field-variable conditions. More importantly, neither study provided explainability of how environmental features relate to visual attention, a gap our Grad-CAM analysis directly addresses.

Compared to XAI studies in agriculture, Bhandari et al. [15] achieved 99.07% accuracy with Grad-CAM explanations, but their model was image-only and could not explain how environmental context influences decisions. Nigar et al. [16] Similarly, applied LIME to image-only models. Our work applies Grad-CAM to a multimodal agricultural model, revealing the causal link between environmental drivers and visual attention.

3.12 Limitations and Future Work

Several limitations should be acknowledged. First, environmental parameters were synthetically generated. While a pilot test on three real images from Mazowe fields showed 100% accuracy, full field validation with physical sensors across multiple growing seasons remains future work. A preliminary comparison of 10 field samples from the 2025 growing season showed conditional sampling distributions closely approximated field measurements (MSV mean temperature: 26.8°C vs. simulated 27.3°C; leafhopper count: 4.9 vs. 5.4; $p>0.05$). Comprehensive field validation is planned using low-cost IoT sensors [13], [14].

Second, the framework focuses on diagnosis and severity estimation. Future work will integrate reinforcement learning for treatment recommendations, creating an end-to-end decision-support system.

Third, expansion to additional diseases and crops would test generalizability. The architecture-agnostic framework could apply to other pathosystems (e.g., wheat rust, cassava mosaic, rice blast) with appropriate parameter adjustments.

Fourth, the framework is presented as a foundational proof-of-concept. Real-world deployment requires validation on edge devices and in resource-constrained settings.

4. Conclusions

This paper presented a multimodal contextual learning framework for maize disease diagnosis that integrates visual symptoms with agro-environmental data. The model achieved a 3.08% absolute accuracy improvement over image-only baselines (94.3% vs. 91.5%, $p<0.001$), reduced MSV-MLN confusion by 40.7%, and provided stage-specific severity estimation with 81.5% accuracy. Grad-CAM analysis revealed that environmental features refine visual attention, increasing focus on lesion regions by 38.1% and localization accuracy by 45.2%, with leafhopper count showing the strongest correlation ($r=0.62$). Ablation studies confirmed environmental features provided the largest contribution (+3.71%, $p<0.001$), while transfer learning proved essential (+9.84%). These results establish that integrating environmental context with visual symptoms enhances diagnostic accuracy and model interpretability. The framework addresses the "agronomic blindness" of image-only models, providing a foundation for context-aware, interpretable

agricultural AI that moves beyond static classification toward actionable decision support

Acknowledgements

This work was supported by the Kemitraan Negara Berkembang (KNB) Scholarship from the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia. The authors declare no conflict of interest.

References

- [1] A. K. Biswal *et al.*, "Maize Lethal Necrosis disease: review of molecular and genetic resistance mechanisms, socio-economic impacts, and mitigation strategies in sub-Saharan Africa," *BMC Plant Biol.*, vol. 22, no. 1, p. 542, Nov. 2022, doi.org/10.1186/s12870-022-03932-y.
- [2] S. Suara, A. Jha, P. Sinha, and A. A. Sekh, "Is Grad-CAM Explainable in Medical Images?," in *Computer Vision and Image Processing*, vol. 2009, H. Kaur, V. Jakhetya, P. Goyal, P. Khanna, B. Raman, and S. Kumar, Eds., in Communications in Computer and Information Science, vol. 2009, Cham: Springer Nature Switzerland, 2024, pp. 124–135. [doi: 10.1007/978-3-031-58181-6_11](https://doi.org/10.1007/978-3-031-58181-6_11).
- [3] T. Ayoub Shaikh, T. Rasool, and F. Rasheed Lone, "Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming," *Comput. Electron. Agric.*, vol. 198, p. 107119, Jul. 2022, [doi: 10.1016/j.compag.2022.107119](https://doi.org/10.1016/j.compag.2022.107119).
- [4] D. Wang, W. Cao, F. Zhang, Z. Li, S. Xu, and X. Wu, "A Review of Deep Learning in Multiscale Agricultural Sensing," *Remote Sens.*, vol. 14, no. 3, p. 559, Jan. 2022, [doi: 10.3390/rs14030559](https://doi.org/10.3390/rs14030559).
- [5] M. Albahar, "A Survey on Deep Learning and Its Impact on Agriculture: Challenges and Opportunities," *Agriculture*, vol. 13, no. 3, p. 540, Feb. 2023, [doi: 10.3390/agriculture13030540](https://doi.org/10.3390/agriculture13030540).
- [6] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, and F. Marinello, "Comparing Inception V3, VGG 16, VGG 19, CNN, and ResNet 50: A Case Study on Early Detection of a Rice Disease," *Agronomy*, vol. 13, no. 6, p. 1633, Jun. 2023, [doi: 10.3390/agronomy13061633](https://doi.org/10.3390/agronomy13061633).
- [7] Md. M. Islam *et al.*, "DeepCrop: Deep learning-based crop disease prediction with web application," *J. Agric. Food Res.*, vol. 14, p. 100764, Dec. 2023, [doi: 10.1016/j.jafr.2023.100764](https://doi.org/10.1016/j.jafr.2023.100764).
- [8] F. Shahoveisi, H. Taheri Gorji, S. Shahabi, S. Hosseiniard, S. Markell, and F. Vasefi, "Application of image processing and transfer learning for the detection of rust disease," *Sci. Rep.*, vol. 13, no. 1, p. 5133, Mar. 2023, [doi: 10.1038/s41598-023-31942-9](https://doi.org/10.1038/s41598-023-31942-9).
- [9] S. A. F. Harahap and I. Irmawan, "Performance comparison of MobileNet, EfficientNet, and Inception for predicting crop disease," *Sriwij. Electr. Comput. Eng. J.*, vol. 1, no. 1, pp. 30–36, Feb. 2024, [doi: 10.62420/selco.v1i1.4](https://doi.org/10.62420/selco.v1i1.4).
- [10] R. R. Patil and S. Kumar, "Rice-Fusion: A Multimodality Data Fusion Framework for Rice Disease Diagnosis," *IEEE Access*, vol. 10, pp. 5207–5222, 2022, [doi: 10.1109/ACCESS.2022.3140815](https://doi.org/10.1109/ACCESS.2022.3140815).
- [11] N. Zhang, H. Wu, H. Zhu, Y. Deng, and X. Han, "Tomato Disease Classification and Identification Method Based on Multimodal Fusion Deep Learning," *Agriculture*, vol. 12, no. 12, p. 2014, Nov. 2022, [doi: 10.3390/agriculture12122014](https://doi.org/10.3390/agriculture12122014).
- [12] J. Su, X. Zhu, S. Li, and W.-H. Chen, "AI meets UAVs: A survey on AI empowered UAV perception systems for precision agriculture," *Neurocomputing*, vol. 518, pp. 242–270, Jan. 2023, [doi: 10.1016/j.neucom.2022.11.020](https://doi.org/10.1016/j.neucom.2022.11.020).
- [13] R. Rashid, W. Aslam, R. Aziz, and G. Aldehim, "An Early and Smart Detection of Corn Plant Leaf Diseases Using IoT and Deep Learning Multi-Models," *IEEE Access*, vol. 12, pp. 23149–23162, 2024, [doi: 10.1109/ACCESS.2024.3357099](https://doi.org/10.1109/ACCESS.2024.3357099).
- [14] O. Elsherbiny, L. Zhou, Y. He, and Z. Qiu, "A novel hybrid deep network for diagnosing water status in wheat crop using IoT-based multimodal data," *Comput. Electron. Agric.*, vol. 203, p. 107453, Dec. 2022, [doi: 10.1016/j.compag.2022.107453](https://doi.org/10.1016/j.compag.2022.107453).
- [15] M. Bhandari, T. B. Shahi, A. Neupane, and K. B. Walsh, "BotanicX-AI: Identification of Tomato Leaf Diseases Using an Explanation-Driven Deep-Learning Model," *J. Imaging*, vol. 9, no. 2, p. 53, Feb. 2023, [doi: 10.3390/jimaging9020053](https://doi.org/10.3390/jimaging9020053).
- [16] N. Nigar, H. Muhammad Faisal, M. Umer, O. Oki, and J. Manappattukunnel Lukose, "Improving Plant Disease Classification With Deep-Learning-Based Prediction Model Using Explainable Artificial Intelligence," *IEEE Access*, vol. 12, pp. 100005–100014, 2024, [doi: 10.1109/ACCESS.2024.3428553](https://doi.org/10.1109/ACCESS.2024.3428553).
- [17] G. Fenu and F. M. Mallocci, "Using Multioutput Learning to Diagnose Plant Disease and Stress Severity," *Complexity*, vol. 2021, no. 1, p. 6663442, Jan. 2021, [doi: 10.1155/2021/6663442](https://doi.org/10.1155/2021/6663442).
- [18] Z.-X. Yang, Y. Li, R.-F. Wang, P. Hu, and W.-H. Su, "Deep Learning in Multimodal Fusion for Sustainable Plant Care: A Comprehensive Review," *Sustainability*, vol. 17, no. 12, p. 5255, Jun. 2025, [doi: 10.3390/su17125255](https://doi.org/10.3390/su17125255).
- [19] L. G. Divyanth, A. Ahmad, and D. Saraswat, "A two-stage deep-learning based segmentation model for crop disease quantification based on corn field imagery," *Smart Agric. Technol.*, vol. 3, p. 100108, Feb. 2023, [doi: 10.1016/j.atech.2022.100108](https://doi.org/10.1016/j.atech.2022.100108).
- [20] Y. Liu *et al.*, "Exploring multi-features in UAV based optical and thermal infrared images to estimate disease severity of wheat powdery mildew," *Comput. Electron. Agric.*, vol. 225, p. 109285, Oct. 2024, [doi: 10.1016/j.compag.2024.109285](https://doi.org/10.1016/j.compag.2024.109285).
- [21] A. S. Keceli, A. Kaya, C. Catal, and B. Tekinerdogan, "Deep learning-based multi-task prediction system for plant disease and species detection," *Ecol. Inform.*, vol. 69, p. 101679, Jul. 2022, [doi: 10.1016/j.ecoinf.2022.101679](https://doi.org/10.1016/j.ecoinf.2022.101679).
- [22] S. Mostafa, D. Mondal, K. Panjvani, L. Kochian, and I. Stavness, "Explainable deep learning in plant phenotyping," *Front. Artif. Intell.*, vol. 6, p. 1203546, Sep. 2023, [doi: 10.3389/frai.2023.1203546](https://doi.org/10.3389/frai.2023.1203546).
- [23] A. Saporta *et al.*, "Benchmarking saliency methods for chest X-ray interpretation," *Nat. Mach. Intell.*, vol. 4, no. 10, pp. 867–878, Oct. 2022, [doi: 10.1038/s42256-022-00536-x](https://doi.org/10.1038/s42256-022-00536-x).
- [24] D. S. Joseph, P. M. Pawar, and K. Chakradeo, "Real-Time Plant Disease Dataset Development and Detection of Plant Disease Using Deep Learning," *IEEE Access*, vol. 12, pp. 16310–16333, 2024, [doi: 10.1109/ACCESS.2024.3358333](https://doi.org/10.1109/ACCESS.2024.3358333).
- [25] G. Dai, J. Fan, Z. Tian, and C. Wang, "PPLC-Net: Neural network-based plant disease identification model supported by weather data augmentation and multi-level attention mechanism," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 5, p. 101555, May 2023, [doi: 10.1016/j.jksuci.2023.101555](https://doi.org/10.1016/j.jksuci.2023.101555).
- [26] Y. Lu, D. Chen, E. Olaniyi, and Y. Huang, "Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review," *Comput. Electron. Agric.*, vol. 200, p. 107208, Sep. 2022, [doi: 10.1016/j.compag.2022.107208](https://doi.org/10.1016/j.compag.2022.107208).
- [27] S. Lamba, A. Baliyan, and V. Kukreja, "GAN based image augmentation for increased CNN performance in Paddy leaf disease classification," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India: IEEE, Apr. 2022, pp. 2054–2059. [doi: 10.1109/ICACITE53722.2022.9823799](https://doi.org/10.1109/ICACITE53722.2022.9823799).
- [28] Y. Haruna, S. Qin, and M. J. Mbyamm Kiki, "An Improved Approach to Detection of Rice Leaf Disease with GAN-Based Data Augmentation Pipeline," *Appl. Sci.*, vol. 13, no. 3, p. 1346, Jan. 2023, [doi: 10.3390/app13031346](https://doi.org/10.3390/app13031346).
- [29] J. Lu, L. Tan, and H. Jiang, "Review on Convolutional Neural Network (CNN) Applied to Plant Leaf Disease Classification," *Agriculture*, vol. 11, no. 8, p. 707, Jul. 2021, [doi: 10.3390/agriculture11080707](https://doi.org/10.3390/agriculture11080707).
- [30] W. Shafik, A. Tufail, C. De Silva Liyanage, and R. A. A. H. M. Apog, "Using transfer learning-based plant disease classification and detection for sustainable agriculture," *BMC Plant Biol.*, vol. 24, no. 1, p. 136, Feb. 2024, [doi: 10.1186/s12870-024-04825-y](https://doi.org/10.1186/s12870-024-04825-y).

- [31] M. Mushayi, H. Shimelis, J. Derera, and S. A. Tesfamariam, "Breeding for resistance to maize streak virus: challenges, progress and future directions: a review," *Front. Plant Sci.*, vol. 16, p. 1590870, Jun. 2025, doi: [10.3389/fpls.2025.1590870](https://doi.org/10.3389/fpls.2025.1590870).
- [32] E. W. Ohlson and J. R. Wilson, "Maize Lethal Necrosis: Impact and Disease Management," *Outlooks Pest Manag.*, vol. 33, no. 2, pp. 45–51, Apr. 2022, doi: [10.1564/v33_apr_02](https://doi.org/10.1564/v33_apr_02).