mIT-CMCA: A Cross-Modal Category Alignment Framework for Robust Maize Disease Identification

Feilong Tang
Universiti of Malaysia Sabah
Rosalyn R Porle
rlyn39@ums.edu.my

Universiti of Malaysia Sabah
Hoe Tung Yew
Universiti of Malaysia Sabah
Farrah Wong
Universiti of Malaysia Sabah

Article

Keywords: Maize disease identification, Multimodal learning, Category-level textual descriptions, Cross-modal alignment, attention mechanism

Posted Date: September 26th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7286789/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on April 21st, 2026. See the published version at https://doi.org/10.1038/s41598-026-48962-w.

mIT-CMCA: A Cross-Modal Category Alignment Framework for Robust Maize Disease Identification

Feilong Tang[1,2], Rosalyn R Porle[1,*], Hoe Tung Yew[1], and Farrah Wong[1]

[1]Faculty of Engineering, Universiti Malaysia Sabah, Kota Kinabalu 88400, Malaysia
[2]College of Mechanical and Electrical Engineering, Xichang University, Xichang 615013, China
[*]rlyn39@ums.edu.my

ABSTRACT

Accurate identification of maize diseases is crucial for safeguarding global food security. Traditional image-based methods often struggle with lighting variations, occlusions, and noise, which limits their robustness and generalisation ability. Multimodal approaches that integrate visual and textual information have shown promise. However, these methods frequently require manually curated textual descriptions for each image, increasing data collection costs and limiting scalability and practical implementation. To address these limitations, we proposed a maize image-text framework with Cross-Modal Category Alignment (mIT-CMCA). This approach enforces category-level alignment between image and text modalities, enabling more accurate and interpretable cross-modal mapping. First, we construct cross-modal representations by aligning image and text modalities at the category level within a shared embedding space. Second, inspired by contrastive learning, we introduce a Cross-Modal Category Alignment (CMCA) loss based on category-level textual descriptions, reducing annotation complexity. Finally, we present an Efficient Channel-Spatial Hybrid Attention (CSHA) module that preserves inter-class boundaries with minimal computational overhead to enhance feature discriminability under complex conditions. Experimental results show that the mIT-CMCA model achieves accuracy, precision, recall, and F1-scores of 99.48%, 99.28%, 99.54%, and 99.41%, respectively, on the maize subset of the PlantVillage dataset (MPVD), representing improvements of 0.24%, 0.13%, 0.17%, and 0.15% over the strongest baseline. On the self-built Maize Leaf-Field dataset (MLFD), the model attains corresponding scores of 93.67%, 93.76%, 93.67%, and 93.66%, with respective improvements of 5.34%, 5.13%, 5.34%, and 5.38%.

***Keywords*—** Maize disease identification, Multimodal learning, Category-level textual descriptions, Cross-modal alignment, attention mechanism

Introduction

Maize annual planting area exceeds 120 million hectares[1]. It is the largest coarse grain crop[2]. However, various maize diseases, such as leaf rust[3] and leaf spot[4], adversely affect maize production by reducing yield, compromising quality, and diminishing economic profitability. As a result, accurate identification of maize leaf diseases is vital for disease management, crop protection, and minimising losses. Traditionally, maize disease identification relies on manual inspections by agricultural experts. However, this method are subjective, labour-intensive, and lack real-time efficiency.

With advanced deep learning, plant disease identification based on images is fast, accurate, and non-destructive[5]. Convolutional Neural Network (CNN)-based approaches, such as GoogleNet[6], ResNet[7], and DenseNet[8], have reduced labour demands and improved the efficiency of disease identification[9–11]. However, their effectiveness in real-world agricultural environments remains limited. Significant intra-class variation, subtle inter-class differences, and various environmental factors still pose substantial challenges[12,13].

To address these challenges, recent models have incorporated attention mechanisms to enhance maize disease identification[14,15]. These models adaptively focus on diseased regions while suppressing background interference. Nevertheless, reliance solely on visual information remains problematic in practice. Occlusion, lighting variability, and subtle symptom expression can obscure critical features and hinder accurate diagnosis. As a result, multimodal approaches have been introduced to integrate visual and textual information, thereby improving robustness and generalisation[16,17]. Additional textual information can reinforce the identification of ambiguous or complex disease symptoms. However, these multimodal approaches often depend on fine-grained, image-level textual annotations. These annotations are costly and limit scalability.

We propose a maize image-text framework with Cross-Modal Category Alignment (mIT-CMCA). The framework projects image and textual features into a shared embedding space. We design a novel Cross-Modal Category Alignment (CMCA) loss. This loss pulls each image feature closer to its corresponding visual synthesis centre and away from others in the shared embedding space. These centres are derived from class-level textual descriptions. We introduce a Channel-Spatial

Hybrid Attention (CSHA) module. This module boosts feature discriminability and preserves inter-class boundaries under complex conditions with minimal computational overhead. Extensive experiments were conducted on the maize subset of the PlantVillage dataset (MPVD), and self-built Maize Leaf-Field dataset (MLFD). The results validated the effectiveness of the proposed mIT-CMCA framework. It consistently outperforms state-of-the-art baselines in performance and generalisation. The main contributions are:

1) This paper proposes a novel framework, mIT-CMCA, for maize leaf disease identification. It achieves significant improvements in classification accuracy and interpretability.

2) This paper introduces a CMCA loss function to enable semantic alignment between image features and category-level textual descriptions.

3) This paper introduces an efficient attention module, CSHA, where efficient channel attention (ECA) channel attention is integrated with spatial attention. The visual backbone's focus on key features is enhanced without a significant increase in model parameters.

4) A self-built Maize Leaf-Field dataset is constructed to assess model robustness under real-world conditions. On this dataset, mIT-CMCA achieves superior performance and demonstrates strong generalisation capability.

5) Comprehensive experiments show the proposed framework outperforms baselines and state-of-the-art methods. It demonstrates strong domain adaptation across both controlled and field environments.

The structure of the paper is as follows: Section 2 overviews related work. Section 3 outlines the materials and methods. Section 4 details the experimental setup and results. Section 5 summarises the key findings and concludes the study.

Related works

Plant diseases profoundly impact crop quality, yield, and productivity. Early diagnosis of these diseases is a significant research focus. Image-based approaches have been widely adopted for plant disease identification. Traditional machine learning methods manually choose features[18,19]. CNNs improve adaptability by learning disease lesion features automatically[20–22]. Dash *et al.*[23] used DenseNet201 with Bayesian optimisation for Support Vector Machines (SVM). Their approach identifies early diseases in maize leaves. Khan *et al.*[24] applied deep transfer learning with models like Visual Geometry Group Network (VGG), Inception V3, ResNet50, and Inception ResNetV2. ResNet50 performed best, achieving an accuracy of 87.51%, a precision of 90.33%, and a recall of 99.80%. Some researchers propose lightweight deep learning models to improve speed and usability. Alpsalaz *et al.*[25] developed an efficient CNN model for maize leaf disease classification. On a four-category maize dataset, it achieved 94.97% accuracy.

Attention mechanisms help models focus on key disease features. Fang *et al.*[26] proposed a neural network using hard coordinated attention (HCA) to extract features from maize leaf disease images at different scales. Their model achieved an average recognition accuracy of 97.75% and an F1 score of 97.03%. Li *et al.*[12] combined multi-dilated modules and Convolutional Block Attention Module (CBAM). On a complex field dataset, this model reached an accuracy of 98.84%. Maizenet[27] uses ResNet-50 with spatial-channel attention. Under various conditions, including different backgrounds, lighting, and variations in colour and size, this method achieved an average accuracy score of 97.89%. Zhu and Gao[28] introduced MC-ShuffLenetV2, adding CBAM to the network for better adaptability and expressiveness. This model achieved 99.86% accuracy on the test set with only 873,936 parameters. Pacal[15] created a lightweight Multi-axis Vision Transformer. This model improved accuracy by replacing convolutional structures in Maxvit with SE blocks. The model reached an accuracy of 99.24%.

Multimodal learning methods have recently gained much attention. Integrating image and textual information brings breakthroughs in plant disease detection. These approaches improve robustness and generalisation by merging complementary data sources. Zhu *et al.*[17] proposed a multimodal model for early potato disease detection. The multimodal MSC-Resvit, MSC-TextCNN, and CT-CNN model achieved 98.43% accuracy on the test set. Multimodal methods have improved model robustness but typically rely on manual text descriptions for each image[16]. This reliance increases data collection costs and limits use in automated agriculture. This study aims to propose a novel image-text multimodal maize disease identification model. The proposed mIT-CMCA uses a shared embedding space and class text descriptions to overcome the limitations of current multimodal approaches. This method further enhances the accuracy and practicality of maize disease identification.

Materials and methods

Datasets

This paper used two datasets: the MPVD and MLFD. These two datasets contain laboratory and field images, respectively. Detailed descriptions and analyses of the two datasets are provided below.

Image acquisition

The PlantVillage[29] dataset has 54,303 leaf images across 38 categories from 14 crop species, such as tomato, maize, and potato. The PlantVillage dataset is a widely used benchmark for automated plant disease diagnosis[30–32]. For this paper, we

selected images from four maize categories—Grey Leaf Spot (GLS), Northern Leaf Blight (NLB), Common Rust (CR), and Healthy—comprising 513, 985, 1,192, and 1,162 images, respectively.

To evaluate the mIT-CMCA in real-world conditions, we curated the MLFD having over 1,000 images captured under natural agricultural conditions. Two plant pathology experts independently annotated the images. Samples with inconsistent cross-validation results or multiple co-occurring diseases were excluded to maintain label reliability. The final dataset has six distinct categories: 78 images of NLB, 77 of GLS, 161 of Phaeosphaeria Leaf Spot (PLS), 63 of CR, 166 of Zinc Deficiency (ZD), and 149 healthy samples. To balance classes, we added public images from PlantVillage, incorporating 144 NLB, 107 CR, and 142 GLS samples. Table 1 presents a summary of the number of laboratory and field images for each class across two datasets.

Dataset	Disease class	Laboratory Images	Field Images
MPVD	GLS	513	0
	NLB	985	0
	CR	1192	0
	Healthy	1162	0
MLFD	NLB	144	78
	GLS	142	77
	PLS	0	161
	Healthy	0	149
	CR	107	63
	ZD	0	166

Table 1. Overview of the maize subset of the PlantVillage dataset (MPVD) and self-built Maize Leaf-Field dataset (MLFD).

Category-level textual descriptions

The proposed mIT-CMCA aligns each image sample and concise category-level textual descriptions in a shared embedding space. Each maize class was assigned a category-level textual description. For example, for the GLS class, the textual description covers the typical colour, shape, and distribution characteristics of the lesions; for the ZD class, the description outlines the typical locations and shapes of chlorotic areas commonly seen in the images. Each class is paired with a concise description, as shown in Figure 1.

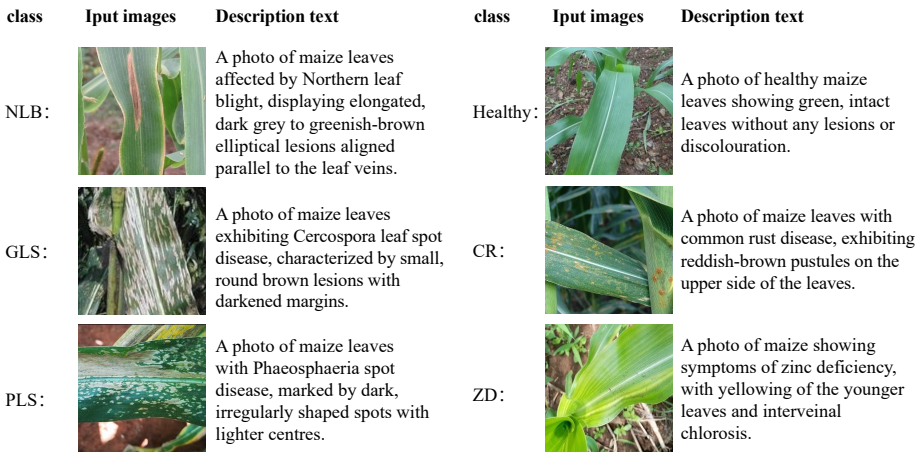

Figure 1. Summary of category-level textual descriptions for maize disease identification.

Data settings

We designed two experimental settings based on dataset size, guided by established data splitting practices to ensure fair evaluation. For the PlantVillage dataset, we used a standard 8:1:1 split for training, validation, and testing[17,33]. In contrast, the Maize Leaf-Field dataset has fewer samples per class. Following the approach for limited data scenarios, we fixed the test set at 50 samples[34,35].

Construction of the mIT-CMCA

Overall structure

The proposed mIT-CMCA performs weak semantic alignment by anchoring image features to category-level textual descriptions. On the textual side, each category-level textual description is processed through a pre-trained text encoder to derive its corresponding visual synthesis centre. On the visual side, an image encoder derives feature representations from the input images. Subsequently, these image feature vectors and visual synthesis centres are projected into a shared embedding space. Cosine similarity guided their alignment. Specifically, this process increases the similarity between each image feature and its corresponding visual synthesis centre, while reducing its similarity to centres of other classes. The alignment process during training is depicted in Figure 2.

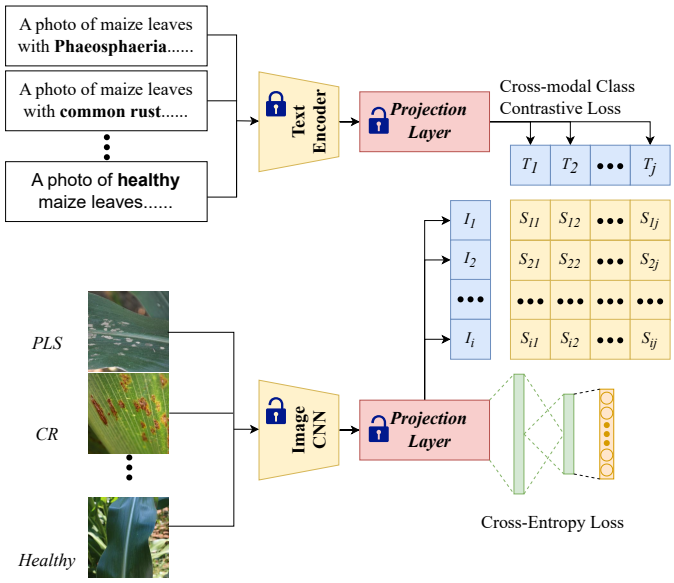

Figure 2. The architecture of the mIT-CMCA framework. The text encoder is frozen during training, while the visual backbone is trainable. The projection layer can be optionally updated, with textual features extracted once or periodically depending on the training mode.

Channel-Spatial Hybrid Attention

Unlike image-level textual descriptions, category-level textual semantics lacks detailed symptom-specific information, lesion location, shape, and colour. The absence of fine-grained cues limits the model's ability to recognise key visual features accurately. To mitigate this limitation, we introduced a novel Channel-Spatial Hybrid Attention (CSHA) module that guides the model to focus on discriminative regions while maintaining low parameter overhead, as illustrated in Figure 3.

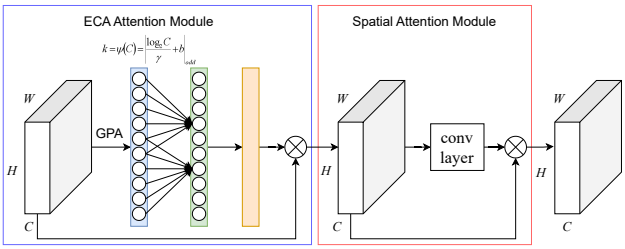

Figure 3. The architecture of the proposed CSHA module. It sequentially applies efficient channel attention (ECA) and spatial attention to refine visual features by emphasising class-relevant regions while suppressing background noise.

The CSHA module integrates ECA with spatial attention. Let the input feature map be denoted as $X \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the number of channels, spatial height, and spatial width, respectively.

In the channel dimension, the input feature map X is first subjected to global average pooling over its spatial dimensions, preserving the number of channels C while reducing the spatial dimensions to 1. The resulting channel descriptor vector $z_{\mathbf{c}} \in \mathbb{R}^C$ is computed as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \tag{1}$$

where z_c vector is processed using a fast 1D convolution with a kernel size of k to compute the channel attention weights efficiently. The channel attention weight map A_C is calculated as follows:

$$A_C = \sigma_1 \left(\text{Conv1D}(z_c); \psi(C) \right) \tag{2}$$

where $\psi(C)$ selects an appropriate kernel size k based on the input channel dimension C. σ_1 is a sigmoid activation function. Specifically, k can be computed as follows:

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + b \right|_{\text{odd}} \tag{3}$$

where γ is a positive scaling coefficient that governs how sensitively k increases with the channel dimension C.

Along the spatial dimension, both max pooling and average pooling are applied to X. The resulting descriptors are then concatenated along the channel axis to form a joint feature representation:

$$M = \text{Concat} \left[M_{\max}(i,j),\ M_{\text{avg}}(i,j) \right] \tag{4}$$

where $M_{\max}(i,j)$ and $M_{\text{avg}}(i,j)$ denote the values obtained by applying max pooling and average pooling at spatial location (i,j), respectively. The spatial attention weights are generated through a two-dimensional (2D) convolution with a 7×7 kernel size and a Sigmoid activation function. The calculation of the spatial attention weight map A_S is as follows:

$$A_S = \sigma_2 \left(\text{Conv2D} \left(M; 7 \times 7 \right) \right) \tag{5}$$

where σ_2 is a sigmoid activation function. Finally, the channel and spatial attention information are fused through element-wise addition, and the resulting feature map Y can be computed as follows:

$$Y = X \otimes A_C \otimes A_S \tag{6}$$

where $\otimes$ denotes element-wise multiplication.

Projection Layer

In the mIT-CMCA, we introduce a Projection Layer in both the image and text branches. Specifically, each category-level text description is semantically encoded, projected to a shared D-dimensional embedding space, and forms category-specific visual synthesis centres. Let $f_I(I_i) \in \mathbb{R}^{d_v}$ and $f_T(T_j) \in \mathbb{R}^{d_t}$ denote the extracted feature representations of the i-th image and the j-th category-level textual description, respectively. The Projection Layer applies learnable linear transformations as follows:

$$z_I = W_I f_I(I_i) + b_I, \quad z_T = W_T f_T(T_j) + b_T \tag{7}$$

where $W_I \in \mathbb{R}^{D \times d_v}$ and $W_T \in \mathbb{R}^{D \times d_t}$ are projection weights and b_I, $b_T \in \mathbb{R}^D$ are the corresponding biases.

These projected features $z_I, z_T \in \mathbb{R}^D$ are semantically aligned and jointly optimised under the proposed contrastive alignment loss. In our implementation, we set $D = 1024$ to balance expressive power and computational cost. The overall structure of the Projection Layer is illustrated in Figure 4.

Cross-modal Category Alignment Loss

The proposed CMCA loss is implemented based on contrastive learning principles. For each image feature vector, the textual description of its ground-truth class is treated as a positive sample, while the descriptions from all others serve as negative samples. Specifically, the CMCA loss maximises the similarity between image features and their corresponding class visual synthesis centres, while minimising their similarity to the centres of other classes, thus enabling effective cross-modal feature matching. This strategy increases inter-class separability and improves the expressiveness of the learned features. Additionally, an intra-batch regularisation term Ω_c was included. The overall CMCA loss, denoted as $\mathscr{L}_{\text{CMCA}}$, is defined as:

$$\mathscr{L}_{CMCA} = -\frac{1}{N} \sum_{i=1}^{N} \left(-\log \frac{\exp\left(\text{sim}\left(c_i^{vc}, c_j^{vsc} \right) / \tau \right)}{\sum_{j=1}^{m} \exp\left(\text{sim}\left(c_i^{vc}, c_j^{vsc} \right) / \tau \right)} \right) + \lambda \Omega_c \tag{8}$$

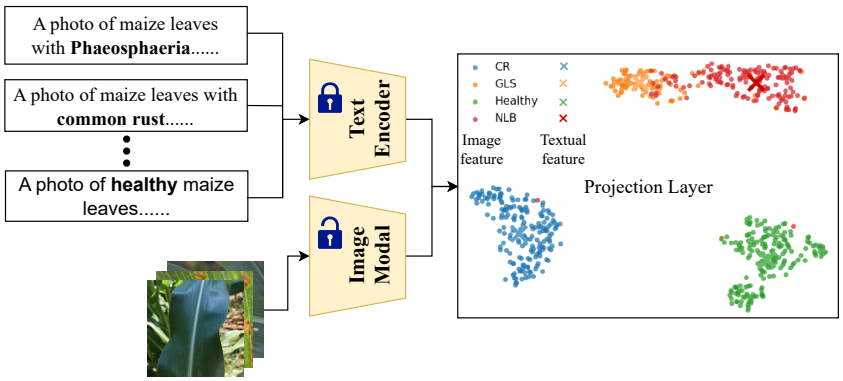

Figure 4. Structure of the Projection Layer. Visual and textual features are projected into a shared embedding space of dimension $D = 1024$ using learnable linear transformations to enable effective cross-modal alignment.

where, m denotes the total number of classes, and N is the batch size. τ is the temperature parameter, typically set to 0.07, as commonly used in contrastive learning and knowledge distillation[36]. c_j^{vsc} denotes the visual synthesis centre of category j, c_i^{vc} denotes the projected visual feature of the i-th image sample, and c_n^{vsc} is the visual synthesis centre corresponding to the ground truth class n for the i-th image sample. The function $\mathrm{sim}(a,b)$ computes the cosine similarity between vectors a and b. λ is a weighting coefficient that controls the influence of the penalty term.

The intra-batch regularisation term Ω_c penalises same-class samples that deviate excessively from their corresponding visual synthesis centres. The penalty term Ω_c is computed as:

$$\Omega_c = \frac{1}{N}\sum_{i=1}^{N} \|c_i^{vc} - c_b^{vsc}\|_2^2 \tag{9}$$

where c_b^{vsc} denotes the batch centre of all samples in the current batch.

To supervise the model's classification capability, we adopt the standard cross-entropy loss $\mathscr{L}_{CE}$, which encourages the model to assign high probability to the correct class label. It is defined as:

$$\mathscr{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n} t_{i,j}\log(p_{i,j}) \tag{10}$$

where N is the batch size, n is the number of classes, $t_{i,j}$ is the one-hot ground-truth label for the i-th sample and class j, and $p_{i,j}$ is the corresponding predicted probability.

We adopt a weighting to balance the contributions of classification supervision and cross-modal alignment. A weighting coefficient α is employed as alignment ratio. Specifically, the total loss can be written as:

$$\mathscr{L}_{total} = \alpha \cdot \mathscr{L}_{CE} + (1-\alpha)\cdot \mathscr{L}_{CMCA} \tag{11}$$

where the alignment ratio $\alpha \in [0,1]$ controls the trade-off between preserving intra-modal key features and enforcing semantic consistency across modalities. $\mathscr{L}_{CE}$ denotes the cross-entropy loss.

Experiments and results

Experimental configuration

The experiments and control experiments were conducted on a server with the following configuration: Intel Core i7-12700H and NVIDIA RTX A2000. We used PyTorch with CUDA 11.7 as the deep learning framework. The batch size for both training and validation was set to 32 due to GPU memory limits. Following the setup used in prior work[35,37], we fixed the number of epochs for all network models at 20. Mixed precision training was enabled to optimise memory usage. AdamW, chosen for its effective weight decay handling and better training stability and generalisation than Adam, was used for model training with a learning rate of 0.0001.

Evaluation indicators

In this paper, we used four common metrics for plant disease identification models: accuracy, precision, recall, and F1 score. These metrics assessed the performance of the maize disease identification model. The calculation formulas are as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{12}$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \tag{13}$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \tag{14}$$

$$\text{F1-score} = \frac{2TP}{2TP+FP+FN} \times 100\% \tag{15}$$

where True Positive (*TP*) is the number of correctly identified positive samples, False Positive (*FP*) is the number of incorrectly identified positive samples, True Negative (*TN*) is the number of correctly identified negative samples, and False Negative (*FN*) is the number of incorrectly identified negative samples.

Results of the mIT-CMCA framework using different models

Experiments on the maize subset of the PlantVillage dataset

The MPVD was used to evaluate the performance of the mIT-CMCA framework, as described in Table 1. The images were resized to a uniform resolution of 224 × 224 pixels. To enhance generalisation, three standard data augmentation techniques were used: random flips, random erasing, and ColourJitter[24,38].

Modular cross-validation experiments were conducted on the MPVD to examine the effect of different image and text encoder combinations. These results, summarised in Table 2, reveal subtle yet meaningful differences in performance across encoder combinations. Overall, the CSHA-DenseNet121 + DistilBERT combination achieved the best results. These results highlight the strong visual feature extraction capability of CSHA-DenseNet121 and the efficient semantic representation of DistilBERT, which together provide a powerful cross-modal fusion mechanism.

Image Encoder	Text Encoder	Accuracy	Precision	Recall	F1
CSHA-GoogLeNet	RoBERTa	99.23	98.80	99.06	98.93
CSHA-GoogLeNet	DistilBERT	99.48	99.07	99.54	99.31
CSHA-GoogLeNet	ELECTRA	98.97	98.77	98.58	98.67
CSHA-ResNet50	RoBERTa	98.97	98.54	98.58	98.56
CSHA-ResNet50	DistilBERT	98.45	98.02	97.62	97.82
CSHA-ResNet50	ELECTRA	97.94	96.92	97.34	97.13
CSHA-DenseNet121	RoBERTa	99.48	99.07	99.54	99.31
CSHA-DenseNet121	DistilBERT	99.48	99.28	99.54	99.41
CSHA-DenseNet121	ELECTRA	99.23	98.64	99.29	98.96

Table 2. The results of different combinations on the MPVD (%). Notably, although the CSHA-GoogLeNet + DistilBERT combination also achieved 99.48% accuracy, its lower F1-score suggests less balanced prediction performance.

To further validate the effectiveness of the proposed model, we compared it against several state-of-the-art methods using the MPVD. As shown in Table 3, the mIT-CMCA model achieved the highest performance across all metrics, with 99.48% accuracy, 99.28% precision, 99.54% recall, and 99.41% F1-score. Compared with MaxViT_tiny, identified as the strongest baseline, mIT-CMCA achieved incremental improvements of 0.24%, 0.13%, 0.17%, and 0.15% in accuracy, precision, recall, and F1-score, respectively. Although these improvements are relatively modest due to the already saturated performance of prior methods, they underscore the effectiveness of incorporating category-level textual descriptions. By introducing cross-modal alignment, mIT-CMCA enhances the model's ability to distinguish between visually similar disease categories.

Experiments on the self-built Maize Leaf-Field dataset

To evaluate model performance in field maize disease identification, we created the MLFD with six maize leaf classes (as shown in Figure 1). All other experimental settings were kept consistent with those used on the MPVD. Modular cross-validation experiments were conducted on the MLFD. The performance of different image and text encoder combinations reveals substantial variation, as shown in Table 4. Among all combinations, CSHA-DenseNet121 + DistilBERT performs best on the MLFD. Experimental results confirm that mIT-CMCA performs robustly in real-world field conditions for maize disease identification.

Model	Accuracy	Precision	Recall	F1
INC-VGGN[39]	91.83	–	–	–
AlexNet_relu2[40]	97.48	96.46	96.52	96.49
RIC-Net[35]	98.45	98.45	98.43	98.43
ClGanNet[41]	98.96	96.10	96.65	96.34
Optimised DenseNet[42]	98.06	97.13	97.38	97.25
LMBRNet[43]	97.25	97.00	97.00	97.00
MaxViT_tiny[15]	99.24	99.15	99.37	99.26
Proposed mIT-CMCA	**99.48**	**99.28**	**99.54**	**99.41**

Table 3. The results of state-of-the-art methods on the maize subset of the MPVD (%).

Image encoder	Text encoder	Accuracy	Precision	Recall	F1
CSHA-GoogLeNet	RoBERTa	90.33	90.66	90.33	90.36
CSHA-GoogLeNet	DistilBERT	92.67	93.48	92.67	92.73
CSHA-GoogLeNet	ELECTRA	91.33	91.78	91.33	91.34
CSHA-ResNet50	RoBERTa	85.67	85.82	85.67	85.60
CSHA-ResNet50	DistilBERT	86.33	86.96	86.33	86.35
CSHA-ResNet50	ELECTRA	86.67	87.19	86.67	86.75
CSHA-DenseNet121	RoBERTa	92.67	92.97	92.67	92.67
CSHA-DenseNet121	DistilBERT	**93.67**	**93.76**	**93.67**	**93.66**
CSHA-DenseNet121	ELECTRA	91.33	91.75	91.33	91.42

Table 4. The results of different modular combinations on the MLFD (%).

We conducted a comparative evaluation on the MLFD to further assess the generalisation capability under real-world conditions. As shown in Table 5, the proposed model achieved the best performance among all evaluated methods. It attained an accuracy, precision, recall, and F1-score of 93.67%, 93.76%, 93.67%, and 93.66%, respectively. Compared with Dy-Tri-ResNet50, identified as the strongest baseline model in the field setting, the mIT-CMCA achieved incremental improvements of 5.34%, 5.13%, 5.34%, and 5.38% in accuracy, precision, recall, and F1-score, respectively. The over 5% performance improvement achieved under field conditions demonstrates the proposed model's superior generalisation ability and robustness. Leveraging category-level textual descriptions effectively distinguishes visually similar diseases in complex agricultural settings.

Model	Accuracy	Precision	Recall	F1
MobileNetV3	75.33	77.33	75.33	75.18
Xception	71.67	72.76	71.67	71.42
ResNet50	83.67	84.69	83.67	83.85
DenseNet121	86.33	87.49	86.33	86.43
InceptionResNetV2	87.67	87.86	87.67	87.67
InceptionV3	88.00	88.53	88.00	88.09
MaizeNet[27]	87.67	87.75	87.67	87.67
Dy-Tri-ResNet50[38]	88.33	88.63	88.33	88.28
Proposed mIT-CMCA	**93.67**	**93.76**	**93.67**	**93.66**

Table 5. Comparison of the results of different models on the MLFD (%).

These findings emphasise the importance of selecting an effective image encoder and a lightweight yet semantically rich text encoder. The consistently strong performance of CSHA-DenseNet121 + DistilBERT on the MPVD and MLFD confirms its robust generalisation capability and suitability for practical deployment in real-world agricultural settings.

Visualisation

This section uses t-SNE to reduce feature dimensions for visualisation. Figure 5 presents the t-SNE plot and confusion matrix of the mIT-CMCA framework using CSHA-DenseNet121 + DistilBERT on the MPVD. In the plot, samples of different maize disease categories form tight, well-separated clusters. There are slight overlaps among GLS, NLB, and CR samples. Specifically, one NLB sample is close to GLS, and one GLS and one CR sample are close to NLB. The confusion matrix shows

one CR sample misclassified as NLB, and one NLB as GLS. These results show effective feature learning and separation. The fully connected layer further optimises results and reduces misclassification.

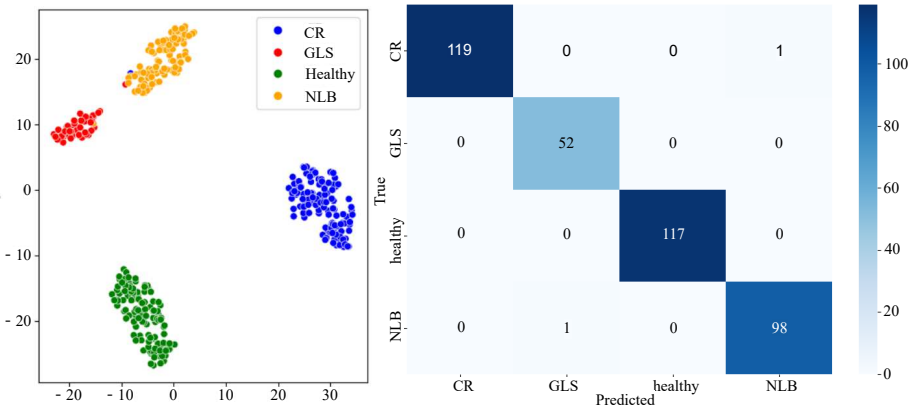

Figure 5. t-SNE visualisation and confusion matrix analysis of the mIT-CMCA method based on the CSHA-DenseNet121 + DistilBERT combination on the MPVD. The class labels are color-coded as follows: CR (blue), GLS (red), Healthy (green), and NLB (yellow).

The t-SNE visualisation and confusion matrix for the mIT-CMCA model (CSHA-DenseNet121 + DistilBERT) on the MLFD are shown in Figure 6. The t-SNE plot displays distinct, compact clusters for different disease classes. A slight overlap exists between GLS and NLB, with some samples highly similar. According to the confusion matrix, most similar samples are correctly classified. For instance, 6 NLB samples resemble GLS in the t-SNE plot, with only 2 NLB misclassified as GLS. This result shows that performance drops slightly with less data or complex backgrounds, but the model separates classes well. Adding a fully connected layer further optimises classification and reduces errors.

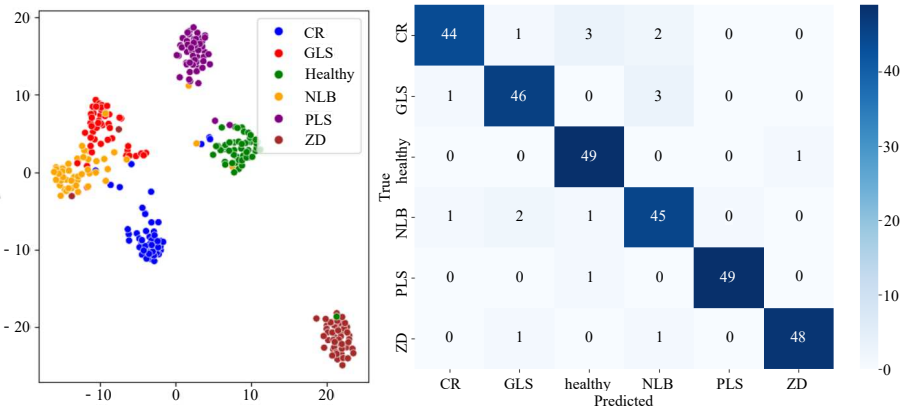

Figure 6. t-SNE Visualisation and Confusion Matrix Analysis of mIT-CMCA based on the CSHA-DenseNet121 + DistilBERT combination on the MLFD. The class labels are color-coded as follows: CR (blue), GLS (red), Healthy (green), NLB (yellow), PLS (purple), and ZD (brown).

Ablation studies

Dimension of Shared Embedding Space D

We first conducted an ablation study on the shared embedding space dimension D by reducing its size step by step. We observed changes in model performance, as shown in Table 6, where D is originally 1024. The dimensionality reduction factor increased from 2^0 to 2^5. The findings show that smaller feature space dimensions cause insufficient semantic expressiveness and decrease performance. Setting the dimension too large introduces redundant information, which weakens the model's focus on key

semantics and lowers performance. After comparison, we found the model performs best when the shared space dimension is set to $D/4$. This setting maintains semantic information and minimises redundant features, striking the best performance balance.

Dimension	D	$D/2$	$D/4$	$D/8$	$D/16$	$D/32$
MPVD Accuracy (%)	98.45	99.23	**99.48**	99.23	98.97	98.20
MLFD Accuracy (%)	91.00	91.67	**93.67**	92.00	92.33	89.00

Table 6. Ablations on varying the dimension of the shared embedding space.

Alignment ratio α

We also conduct an ablation study on the loss function's weighting coefficient α. The value of α affects the model's balance between distinguishing categories and aligning cross-modal features. We perform ablation experiments with different α values on the MPVD and MLFD. This study helps determine the best loss weighting strategy. When $\alpha = 0$, the model excludes the cross-entropy loss and relies only on the CMCA loss for optimisation. Predictions then depend solely on image and text feature similarity. When $\alpha = 1.0$, the classification is based only on the image classifier and ignores cross-modal constraints. Table 7 presents the results of the ablation study on alignment ratio α. We find the optimal α is 0.4. This analysis helps identify the best balance between classification accuracy and feature alignment.

α	0	0.2	0.4	0.6	0.8	1.0
MPVD Accuracy (%)	97.94	99.23	**99.48**	99.23	98.97	98.71
MLFD Accuracy (%)	90.33	92.00	**93.67**	92.33	92.00	90.67

Table 7. Ablations on varying the weight α of the image-text alignment term.

Conclusion

This paper introduces image-text multimodal learning for maize disease identification using category-level textual descriptions. This approach greatly reduces the annotation cost usually required for multimodal training data. It allows existing image-only datasets to be quickly adapted to the proposed multimodal framework, removing the need for labor-intensive image-level annotations. The mIT-CMCA model shows strong generalisation, achieving 99.48% accuracy on the MPVD and 93.67% on the MLFD. It outperforms state-of-the-art models on the PlantVillage benchmark, highlighting its strong potential for real-world deployment. The performance of mIT-CMCA is influenced by two critical hyperparameters: the shared embedding space dimension D and the alignment ratio α, both of which exhibit non-monotonic effects on accuracy. Model accuracy improves as projection dimension D increases up to 256, but declines if increased further. Varying α reveals the best performance at α=0.4. This work makes a valuable contribution by exploring ways to lower the data acquisition burden for image-text multimodal methods. It offers practical guidance for applying such approaches in fields with limited or no image-level textual supervision.

Author contributions statement

Feilong Tang conceived and designed the experiments, collected the dataset, and implemented the proposed model. Feilong Tang and Hoe Tung Yew conducted the experiments and performance evaluations. Rosalyn R Porle supervised the study and provided guidance on model design and manuscript structure. Farrah Wong and Hoe Tung Yew assisted in data analysis, result interpretation, and visualisation. All authors reviewed and approved the final manuscript.

Additional information

Correspondence and requests for materials should be addressed to F.T. or R.P.

Declaration of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wang, H., Pan, X., Zhu, Y., Li, S. & Zhu, R. Maize leaf disease recognition based on tc-mrsn model in sustainable agriculture. *Comput. Electron. Agric.* **221**, 108915 (2024).
2. International Grains Council. IGC – International Grains Council. https://www.igc.int/en/default.aspx (2025). Accessed: 15 May 2025.
3. Wang, P. *et al.* Efficient and accurate identification of maize rust disease using deep learning model. *Front. Plant Sci.* **15**, 1490026 (2025).
4. Yang, S. *et al.* Classification and localization of maize leaf spot disease based on weakly supervised learning. *Front. Plant Sci.* **14**, 1128399 (2023).
5. Bachhal, P., Kukreja, V. & Ahuja, S. Maize disease classification using deep learning techniques: a review. In *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 259–264 (IEEE, 2023).
6. Szegedy, C. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).
7. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
8. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
9. Wang, Y., Wang, H. & Peng, Z. Rice diseases detection and classification using attention based neural network and bayesian optimization. *Expert. Syst. with Appl.* **178**, 114770 (2021).
10. Zhang, Y., Huang, S., Zhou, G., Hu, Y. & Li, L. Identification of tomato leaf diseases based on multi-channel automatic orientation recurrent attention network. *Comput. Electron. Agric.* **205**, 107605 (2023).
11. Lanjewar, M. G. & Panchbhai, K. G. Convolutional neural network based tea leaf disease prediction system on smart phone using paas cloud. *Neural Comput. Appl.* **35**, 2755–2771 (2023).
12. Li, E. *et al.* A novel deep learning method for maize disease identification based on small sample-size and complex background datasets. *Ecol. Informatics* **75**, 102011 (2023).
13. Wu, X. *et al.* From laboratory to field: Unsupervised domain adaptation for plant disease recognition in the wild. *Plant Phenomics* **5**, 0038 (2023).
14. Yang, L. *et al.* Googlenet based on residual network and attention mechanism identification of rice leaf diseases. *Comput. Electron. Agric.* **204**, 107543 (2023).
15. Pacal, I. Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model. *Expert. Syst. with Appl.* **238**, 122099 (2024).
16. Zhou, J. *et al.* Crop disease identification and interpretation method based on multimodal deep learning. *Comput. Electron. Agric.* **189**, 106408 (2021).
17. Zhu, H. *et al.* Potato disease detection and prevention using multimodal ai and large language model. *Comput. Electron. Agric.* **229**, 109824 (2025).
18. Devi, K. S., Srinivasan, P. & Bandhopadhyay, S. H2k–a robust and optimum approach for detection and classification of groundnut leaf diseases. *Comput. Electron. Agric.* **178**, 105749 (2020).
19. Sun, Y., Jiang, Z., Zhang, L., Dong, W. & Rao, Y. Slic_svm based leaf diseases saliency map extraction of tea plant. *Comput. electronics agriculture* **157**, 102–109 (2019).
20. Abd Algani, Y. M. *et al.* Leaf disease identification and classification using optimized deep learning. *Meas. Sensors* **25**, 100643 (2023).
21. Faisal, M., Leu, J.-S. & Darmawan, J. T. Model selection of hybrid feature fusion for coffee leaf disease classification. *IEEE Access* **11**, 62281–62291 (2023).
22. Guerrero-Ibañez, A. & Reyes-Muñoz, A. Monitoring tomato leaf disease through convolutional neural networks. *Electronics* **12**, 229 (2023).
23. Dash, A., Sethy, P. K. & Behera, S. K. Maize disease identification based on optimized support vector machine using deep feature of densenet201. *J. Agric. Food Res.* **14**, 100824 (2023).

24. Khan, I., Sohail, S. S., Madsen, D. Ø. & Khare, B. K. Deep transfer learning for fine-grained maize leaf disease classification. *J. Agric. Food Res.* **16**, 101148 (2024).

25. Alpsalaz, F., Özüpak, Y., Aslan, E. & Uzel, H. Classification of maize leaf diseases with deep learning: Performance evaluation of the proposed model and use of explicable artificial intelligence. *Chemom. Intell. Lab. Syst.* 105412 (2025).

26. Fang, S. *et al.* Multi-channel feature fusion networks with hard coordinate attention mechanism for maize disease identification under complex backgrounds. *Comput. Electron. Agric.* **203**, 107486 (2022).

27. Masood, M. *et al.* Maizenet: A deep learning approach for effective recognition of maize plant leaf diseases. *IEEE Access* **11**, 52862–52876 (2023).

28. Zhu, S. & Gao, H. Mc-shufflenetv2: A lightweight model for maize disease recognition. *Egypt. Informatics J.* **27**, 100503 (2024).

29. Mohanty, S. P., Hughes, D. P. & Salathé, M. Using deep learning for image-based plant disease detection. *Front. plant science* **7**, 215232 (2016).

30. Ashmafee, M. H. *et al.* An efficient transfer learning-based approach for apple leaf disease classification. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–6 (IEEE, 2023).

31. Ouamane, A. *et al.* Enhancing plant disease detection: A novel cnn-based approach with tensor subspace learning and howsvd-mda. *Neural Comput. Appl.* **36**, 22957–22981 (2024).

32. Roumeliotis, K. I., Sapkota, R., Karkee, M., Tselikas, N. D. & Nasiopoulos, D. K. Plant disease detection through multimodal large language models and convolutional neural networks. *arXiv preprint arXiv:2504.20419* (2025).

33. Xiong, Y., Liang, L., Wang, L., She, J. & Wu, M. Identification of cash crop diseases using automatic image segmentation algorithm and deep learning with expanded dataset. *Comput. Electron. Agric.* **177**, 105712 (2020).

34. Kumar, A., Nelson, L. & Venu, V. S. Efficientnet-bl based maize plant leaf disease classification using deep learning. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 1636–1642 (IEEE, 2024).

35. Zhao, Y., Sun, C., Xu, X. & Chen, J. Ric-net: A plant disease classification model based on the fusion of inception and residual structure and embedded attention mechanism. *computers Electron. Agric.* **193**, 106644 (2022).

36. Zhang, S. *et al.* Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3479–3488 (2023).

37. Chi, Z. *et al.* Learning to adapt frozen clip for few-shot test-time domain adaptation. *arXiv preprint arXiv:2506.17307* (2025).

38. Tang, F., Porle, R. R., Yew, H. T. & Wong, F. Identification of maize diseases based on dynamic convolution and tri-attention mechanism. *IEEE Access* (2025).

39. Chen, J., Chen, J., Zhang, D., Sun, Y. & Nanehkaran, Y. A. Using deep transfer learning for image-based plant disease identification. *Comput. electronics agriculture* **173**, 105393 (2020).

40. Barburiceanu, S., Meza, S., Orza, B., Malutan, R. & Terebes, R. Convolutional neural networks for texture feature extraction. applications to leaf disease classification in precision agriculture. *IEEE Access* **9**, 160085–160103 (2021).

41. Sharma, V., Tripathi, A. K., Daga, P., Mittal, H. *et al.* Clgannet: A novel method for maize leaf disease identification using clgan and deep cnn. *Signal Process. Image Commun.* **120**, 117074 (2024).

42. Waheed, A. *et al.* An optimized dense convolutional neural network model for disease recognition and classification in corn leaf. *Comput. Electron. Agric.* **175**, 105456 (2020).

43. Li, M., Zhou, G., Chen, A., Li, L. & Hu, Y. Identification of tomato leaf diseases based on lmbrnet. *Eng. Appl. Artif. Intell.* **123**, 106195 (2023).