Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A TRUSTWORTHY MULTIMODAL FOUNDATION FRAMEWORK WITH ADAPTIVE CROSS-MODAL ATTENTION FUSION FOR OPEN-WORLD TOMATO DISEASE DIAGNOSIS IN REAL AGRICULTURAL ENVIRONMENTS

RAJI NETTATH[1], DR. VIJAYABHANU RAJAGOPAL[2]

[1] Research Scholar, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore
[2] Associate Professor, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore
Corresponding Author Email: nettathraji@gmail.com

ABSTRACT

Diseases are important factors which threaten the productivity of tomato crops worldwide and result in heavy yield losses, monetary losses to other crops and pose serious challenges. Despite the recent progress of automated diagnosis of plant diseases using deep learning methods using only RGB images, most of the current deep learning approaches are only based on RGB images, require closed-set environments, do not account for environmental variability, and are unable to account for prediction reliability in real-world field environments or to deal with unknown diseases. Considering the above issues, the present paper introduces a novel Trustworthy Multimodal Foundation Framework (TMFF) for open world tomato disease diagnosis for precision agriculture. The proposed framework includes a pre-trained vision foundation model and various heterogeneous environmental factors like weather data, soil characteristics, crop growth stage, and Internet of Things (IoT) sensor data to learn powerful multimodal representations of diseases. Propose an adaptive cross-modal attention fusion module (ACMAF) that can adaptively learn the complementary interactions between the visual and environmental features through bidirectional attention and adaptive modality weighting. Also, an Open-World Recognition Module (OWRM) is capable of reliably recognizing known and unseen disease categories based on prototype-based representation learning and adaptive threshold estimation. In order to improve trust in decision-making, a Trustworthiness Assessment Module (TAM) will have Bayesian uncertainty estimation, calibrated confidence and explainable artificial intelligence-based prediction. Experimental evaluation shows that TMFF's classification achieves 99.18%, F1-score achieves 99.09%, macro-AUC achieves 0.9987 and unknown disease detection achieves 97.88%, which is better than state-of-the-art convolutional neural networks, vision transformers and foundation models while allowing real-time inference. According to this proposal, an integrated reliable and accurate tomato disease diagnosis solution can be developed. This will have great potential applications for precision agriculture and next-generation artificial intelligence-enabled crop health monitoring.

Keywords: *Tomato Disease Diagnosis; Foundation Models; Multimodal Learning; Adaptive Cross-Modal Attention Fusion; Open-World Recognition; Trustworthy Artificial Intelligence; Vision Transformers; Precision Agriculture; Explainable Artificial Intelligence; Edge Intelligence.*

1. INTRODUCTION

Tomato (Solanum lycopersicum L.) is one of the most grown horticultural crops worldwide. It plays a significant role in food security, human nutrition and human economy at global level. Tomato is a fresh and processed food with high content of vitamins, minerals and antioxidants as well as dietary fibre. Furthermore, tomatoes are one of the significant crops in contemporary agriculture. FAO reports that the world production of tomatoes is growing because of growing consumer demand and better new technologies for better cultivation [40].

Nevertheless, tomato production is highly susceptible to an array of fungal, bacterial, viral, and physiological diseases which greatly reduce yield, fruit quality and farmers' income. Early blight, late blight, bacterial spot, leaf mold, septoria leaf spot, target spot, tomato mosaic virus,

and tomato yellow leaf curl virus are the most serious diseases that can cause high losses under favourable conditions if not diagnosed and controlled early enough [19].

Foliar symptom assessment by qualified plant pathologists or agricultural specialists, is the traditional method of tomato disease diagnosis. Manual diagnosis can achieve good accuracy in a controlled environment, but is labour-intensive, subjective and impractical at large commercial farms. This means that the use of excessive pesticides, environmental damage, and loss of economic value are all possible due to delayed diagnosis that stems from clinical symptoms similar to other diseases, varying climatic conditions and lack of expert knowledge [3, 19], making intelligent, automated and reliable disease diagnosis systems an important research field in precision agriculture.

1.1 Evolution of Artificial Intelligence for Tomato Disease Diagnosis

With the recent advancement in AI, computer vision and deep learning, automated diagnosis of plant disease is possible. Previous architectures based on handcrafted image descriptors, such as colour histogram, texture feature, shape-based representations, followed by a normal classifier like SVM, Random Forest, Artificial Neural Network, etc. These methods achieved moderate success but were heavily reliant on manual feature engineering, and their robustness suffered in complications in the agricultural domain.

The advent of convolution neural networks (CNNs) improved disease analysis because of the high-end feature learning from images. According to previous studies [1]-[4], plant disease classification methods based on architectures like AlexNet, VGGNet, ResNet, DenseNet, EfficientNet, and Inception outperform traditional machine learning methods. As indicated in [2], CNNs based models have achieved more than 95% classification accuracy using PlantVillage benchmark dataset.

Thus, deep learning has become the trending solution in the agricultural image domain. Later works extended the diagnosis of diseases to the use of object detection frameworks, such as Faster R-CNN, SSD, YOLOv3, YOLOv5, and YOLOv8, to localize and identify disease lesions in complex scenes [5]–[7], [26]–[29].

Recently, transformer-type architectures like Vision Transformer (ViT) and Swin Transformer can model long-range spatial relationships better through self-attention mechanisms [21], [22]. In parallel, results showcased by foundation vision models, such as DINOv2, CLIP, and SigLIP, have demonstrated outstanding transfer learning performance. Large-scale self-supervised pretraining can be largely attributed to representation learning and generalisation with these models [23]–[25].

Technological advancements have significantly improved smart crop disease diagnosis. However, the problem remains that scaling these outcomes to robust field-ready agricultural systems is difficult.

1.2 Limitations of Existing Disease Diagnosis Systems

Even with significant advancements, several inherent limitations continue to constrain the practical application of existing systems for the diagnosis of disease in tomato. To begin with, the bulk of the existing studies are trained and evaluated on laboratory-controlled datasets, PlantVillage, where leaf images are taken under a uniform background, controlled illumination and insignificant environmental variations [3], [4]. When deployed in real agricultural environments with varied lighting, shadows, occlusions, overlapping leaves, multiple disease symptoms, and cluttered backgrounds, these models trained on these datasets often show huge performance drop [5], [17], [18].

In the second place, most of the approaches model disease diagnosis as a closed-set classification problem, which assumed that every testing sample belongs to one of the disease categories that the model observed during training. In practical agriculture, it is often not true that we will not see newly appearing pathogens, nutrient deficiencies, abiotic stresses, mixed infections and unknown disease symptoms. In such cases, traditional deep learning frameworks could provide sufficiently confident yet wrong predictions, thus leading to higher chances of wrong decision making for disease management [30]−[32].

In addition, most of the existing disease diagnosis systems depend on RBG images and do not consider very influential environmental factors. Plant diseases depend on various factors. The factors include environmental, biological and plant factors. In addition, the diseases infect plants under optimum conditions and weather.

There have been a few studies that investigate multimodal learning through image data and environment information. However, it is mostly feature concatenation based and does not include any complex interactions between heterogeneous modalities [9]. Currently, most deep learning

approaches are effectively black-box; that is, given some input a model can typically predict a disease. Nevertheless, the model does not indicate how confident it is of this prediction, or why it reached this conclusion. In precision agriculture, mispredictions but with high confidence can delay treatment, spray pesticides unnecessarily and cause high production costs. Trustworthy developments of artificial intelligence like explainability, uncertainty estimation and confidence calibration are helping reliable agricultural decision making [33]–[36].

Eventually, many existing frameworks take into account the classification accuracy but do not take into consideration the practical deployment aspects, like computational efficiency, robustness against distribution shifts, and compatibility with edge devices typically common in smart farming [10], [17], [37]–[39].

1.3 Proposed Trustworthy Multimodal Foundation Framework (TMFF)

TMFF integrates a pretrained Vision Foundation Encoder with environmental information using Adaptive Cross-Modal Attention Fusion (ACMAF). The main idea of TMFF is to combine a Vision Foundation Encoder to environmental information through Adaptive Cross-Modal Attention Fusion (ACMAF) in a pre-trained manner. In addition, it includes Open-World Recognition (OWRM) and Trustworthiness Assessment (TAM) to enable unknown disease detection, uncertainty estimation, confidence calibration, and explainability.

1.4 Main Contributions

The proposed study brings together a multimodal framework, adaptable fusion process, open world recognition ability and trustworthy prediction module, which can be used for robust tomato disease diagnosis in actual agricultural setting with an extensive evaluation protocol. The rest of the paper is sequenced as follows. Section 2 presents a literature review on the diagnosis of tomato diseases, foundation models, multimodality, open-world recognition, and trustworthy AI. The proposed architecture of the TMFF and its constituent modules is discussed in section 3. In Section 4, the mathematical structure of the framework is established. Experimental setup, datasets and evaluation methodology are described in section 5. The experimental results will be presented in Section 6, The session 7 describes the discussion and to conclude, Section 8 wraps up the paper and proposes future research directions.

2. RELATED WORK

The rapid development of artificial intelligence (AI), computer vision and precision agriculture has moved automated tomato disease diagnosis from image classification to an intelligent decision support paradigm to support sustainable crop production. In the early days, image classification often used handcrafted image descriptors and standard machine learning algorithms. In this process, features like color, texture and shape were manually extracted and then classification was performed. Despite demonstrating acceptable performance under laboratory conditions, they were strongly affected by the illumination, background complexity, and severity of diseases. The introduction of deep learning, especially CNNs, helped to automatically extract hierarchical features from images, which resulted in a boom in deep learning. Ever since that, more research on far more accurate and deployable disease diagnosis system is going on with the help of transformer architectures, multimodal learning, foundation models, XAI, edge etc [1]–[9]. These advances notwithstanding, critical challenges remain about the robustness and generalization capacities of these models to recognize unknown diseases with reliable confidence in real agricultural conditions. The limitations mentioned above motivate the development of the proposed Trustworthy Multimodal Foundation Framework (TMFF) with Multi-modal Representation Learning, Open-world Recognition, and Trustworthy AI as building blocks.

CNN-based techniques now dominate tomato disease diagnosis because they learn hierarchical visual representations automatically, negating the need for tailored feature engineering.

Initial research using AlexNet and GoogLeNet on benchmark datasets such as PlantVillage revealed that transfer learning architecture can deliver classification accuracies of greater than 99% [2], [3]. After that, several architectures were created like ResNet, DenseNet, EfficientNet, Inception, MobileNet, and attention-enhanced residual networks, which greatly improved feature extraction and disease discrimination [8], [11]–[16], [20]. Also, multi-scale feature learning and attention mechanisms were used to identify visually similar disease symptoms better [12], [13] Nevertheless, the majority of CNN-based research was conducted on laboratory-controlled datasets with isolated leaves, homogeneous background and uniform light [3], [4]. As a result, these models exhibit significant performance loss when deployed in actual agricultural environments with

cluttered backgrounds, changing weather conditions, occlusion of crops by foliage, shadows, and varying lighting [17], [18].

Importantly, CNNs are powerful feature extractors; however, they only learn from image data and assume a closed set, limiting practical use in precision agriculture.

Disease diagnosis has been further extended with the use of object detection networks that can classify disease types and localize the infected area on plant images. Frameworks such as faster R-CNN, SSD and more recently the YOLO family have made real-time suitable lesion detection possible which are applicable for greenhouse monitoring and in-field applications. The updated versions of YOLOv5 and YOLOv8 have shown an exceptional detection accuracy with the computational efficiency demanded by real-time deployment. However, these methods mostly rely on supervised closed-set learning, assuming that each testing sample comes from a disease class appearing in training. They mostly disregard environmental factors that impact disease incidence and progression. They also have limited capabilities at managing unseen diseases and estimating uncertainties in prediction. As a result, their robustness deteriorates with novel pathogens, mixed infections, and rapidly changing agricultural situations.

In visual representation learning, the introduction of Vision Transformers (ViTs) and foundation vision models has transformed the dynamics. It has been suggested in [21], [22] that self-attention mechanisms trump CNNs for capturing long-range spatial dependencies and global context. Hierarchical transformer structures, like Swin Transformer, enhance computational efficiency to accommodate large-sized images from agriculture. Moreover, self-supervised foundation models (i.e., DINOv2) and vision-language models (i.e. CLIP) provide highly transferable visual representations learned from large datasets without extensive manual annotations. These models demonstrate stronger generalization across different contexts, as well as less reliance on large amounts of labeled agricultural data. However, today's tomato disease diagnosis applications are primarily focused on images and hardly utilize contextual agricultural information, which prevents them from capturing the complex interactions between the environment and disease. In addition, due to the absence of uncertainty estimation, explainability and open-world recognition mechanisms, foundation models cannot be deployed in safety-critical agricultural decision-support systems.

The incidence of plant diseases is linked not only to symptoms visible on the plants themselves but also to environmental factors such as temperature, humidity, rainfall, soil moisture, nutrient levels, irrigation and crop stage. As a result, multimodal education has become a promising approach in precision agriculture [9], [10]. Recent studies show that combining environmental measurements with RGB images improves classification accuracy over using images alone [9]. The recent advances in Internet of Things (IoT) technologies allow the continuous acquisition of environmental information by employing a distributed sensor network. They enable real-time crop monitoring and intelligent farming systems [37], [39]. Most existing multimodal methods employ basic feature concatenation or shallow fusion techniques and are unable to draw complex relations among heterogeneous modalities. The lack of proposed mechanisms for dynamic interaction adaptive weights over visual and environmental information is notably an open issue. TMFF proposes an Adaptive Cross-Modal Attention Fusion module addressing this.

A key deficiency associated with existing disease diagnosis systems is closed-set classification. Agricultural settings often harbor pathogens that were never encountered during training, along with novel diseases, nutrient deficiencies, abiotic stresses and mixed infections. The issue of the unknown unknowns is solved by open world recognition in which models differentiate known diseases from anything unknown [30]–[32]. In the same manner, trustable artificial intelligence has gained prominence as deep learning models often act as black boxes providing high-confidence predictions without relaying any uncertainty or explaining their reasoning. Recent developments in Bayesian uncertainty estimation, confidence calibration and explainable AI techniques, such as Grad-CAM and SHAP have enhanced the transparency and reliability of the developed models [33]–[36]. However, such capabilities are seldom integrated into the existing frameworks of tomato disease diagnosis. All together, these limitations highlight the necessity of a unified architecture for integrating foundation vision models, multimodal world learning, adaptive cross-modal fusion, open-world recognition, uncertainty estimation, confidence calibration, and explainable AI. The TMFF proposed will fulfil these research gaps for intelligent tomato disease diagnosis in real agricultural environments in a comprehensive, reliable and credible way.

2. LITERATURE REVIEW

2.1 Research Gap

To begin with, most of the existing studies are heavily dependent on datasets under laboratory control and therefore have limited robustness under real agricultural settings [17], [18]. In addition, the existing disease diagnosis systems depend heavily on visual features and do not use excessive amount of environmental, soil, crop and IoT sensors' data which actually plays a very crucial role in disease development [9], [37]. Also, traditional classifiers operate while assuming closed-world learning and can't detect new disease types. Lastly, most existing models focus on classification accuracy and fail to consider other features like explainability, uncertainty estimation.

Table 1. Comparison of representative tomato disease diagnosis approaches.

Approach	Real Field	Multimodal	Foundation Model	Open-World	Trustworthy AI
CNN [1]–[3]	✗	✗	✗	✗	✗
Faster R-CNN [5], [6]	Partial	✗	✗	✗	✗
YOLO [7], [26]–[29]	Partial	✗	✗	✗	✗
Vision Transformer [21], [22]	Partial	✗	Partial	✗	Partial
Foundation Models [23]–[25]	✓	✗	✓	✗	Partial
Multimodal Learning [9]	Partial	✓	✗	✗	✗
Proposed TMFF	✓	✓	✓	✓	✓

The review suggests that presently there does not exist a single framework that brings together foundation vision models, adaptive multimodal fusion, open-world recognition, and trustworthy AI in a unified architecture for tomato disease diagnosis. To overcome such limitations, we next propose a Trustworthy Multimodal Foundation Framework (TMFF), consisting of a Vision Foundation Encoder, an Adaptive Cross-Modal Attention Fusion (ACMAF) module, an Open-World Recognition Module (OWRM) and a Trustworthiness Assessment Module (TAM) for reliable diagnosis in real agricultural environments.

3. PROPOSED FRAMEWORK

The increasing complexity of tomato disease diagnosis in the real agricultural environment has generated a demand for the intelligent system that can efficiently integrate heterogeneous information while providing a prediction that is accurate, robust, and reliable. In spite of the recent advances in deep learning for automated disease recognition, most existing approaches face several practical limitations. To begin with, they rely primarily on the RGB images taken at controlled lab conditions and thus don't generalize well to real agricultural settings featuring varying illumination, confusing backgrounds, occlusion, multiple disease symptoms and environment variations. The second point is the majority of existing methods work under the closed-set assumption in which every testing sample is meant to belong to one of the disease categories seen during training. Yet, the real world of farming is filled with novel pathogens, co-infections, nutrient deficits, abiotic stresses, and unknown disease situations not represented sufficiently by these classes.

The normal deep learning models are often black-box systems that do not provide uncertainty estimation or interpretable explanations for their highly confident predictions. Consequently, this limits user trust and practical usability in precision agriculture [9], [17], [18], [21]– [36]. The drawbacks above inspire architects to design a unified framework to bring multimodal representation, open-world recognition and trustworthy AI together in an end-to-end method.

To fix these issues, this paper presents the Trustworthy Multimodal Foundation Framework (TMFF), a practical Open-World Tomato Disease Diagnosis framework to be used in real-world agricultural settings. TMFF leverages a multitude of heterogeneous information sources such as RGB leaf images, environmental observations, soil characteristics, crop growth information, and IoT sensor information, unlike image-based systems which work on individual leaf image-based systems. The proposed framework leverages the latest advances in foundation vision models, transformer-based multimodal learning, adaptive attention mechanism, open-world recognition, uncertainty estimation, confidence calibration and explainable AI in a united learning framework. Instead of just foreseeing a disease label, TMFF

Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

offers trustworthy diagnostic information with confidence estimation, uncertainty quantification, unknown disease identification, and visual explanations for practical agricultural decisions. The proposed framework's overall architecture is shown in figure 1.

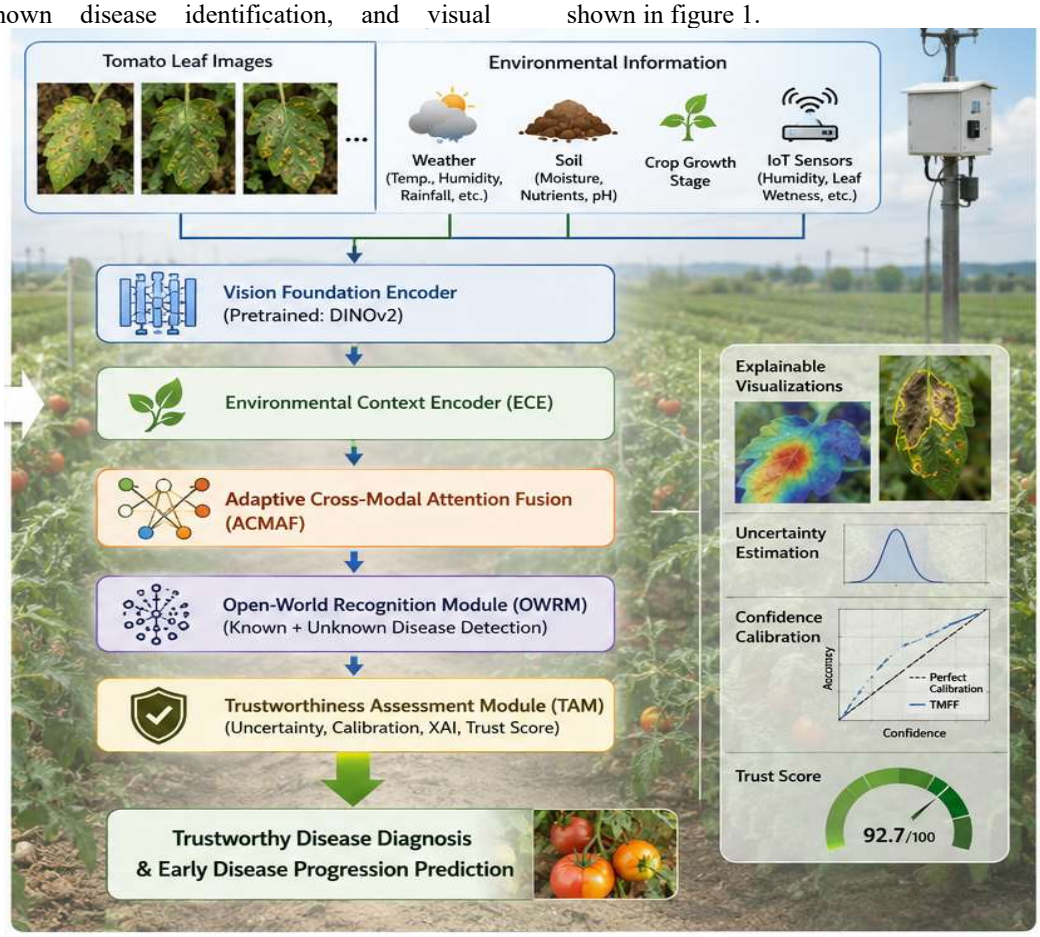


Figure 1. Overall architecture of the proposed Trustworthy Multimodal Foundation Framework (TMFF)

The TMFF (Trustworthy Multimodal Foundation Framework) proposed framework incorporates six different modules which work together to perform robust and reliable tomato disease diagnosis. The VFE is a pretrained transformer-based foundation model that extracts visual features from the tomato's leaf image and uses them as an input feature. The encoder generates strong features under different illumination, background complexity and occlusion conditions through self-attention by modelling local lesion characteristics and global contextual relationships [21]–[25]. Simultaneously, the Environmental Context Encoder (ECE) analyses different farm and agricultural related information like weather, soil properties, crop stage, IoT sensors' readings to generate embeddings that are contextual to provide complementary features to the images for enhancing disease discrimination in various environments [9],[19]. The principal innovation of TMFF is the proposed Adaptive Cross-Modal Attention Fusion (ACMAF) module, which integrates extracted representations. In contrast to the standard feature concatenation methods, our ACMAF uses bidirectional cross-attention and adaptive modality weights to facilitate dynamic learning of complementary interactions between the visual symptoms and environmental context, thus yielding informative multimodal representations for disease classification.

The Open-World Recognition Module (OWRM) analyzes the fused features that merge confidence estimation and uncertainty-aware decision making. This approach helps in distinguishing known diseases from unseen disease conditions, thus reducing misclassifications in real-life agricultural environments [30]–[32]. To improve the reliability of predictions, the Trustworthiness Assessment Module (TAM) incorporates explainability, uncertainty estimation, and confidence calibration

[33]-[35] to make the diagnostic decisions transparent and trustworthy. The outputs of all preceding modules in the Decision Layer are combined to get the predicted class of the disease along with a confidence score, uncertainty estimate, open-world recognition flag, and visual explanation for accurate and interpretable diagnosis of tomato diseases.

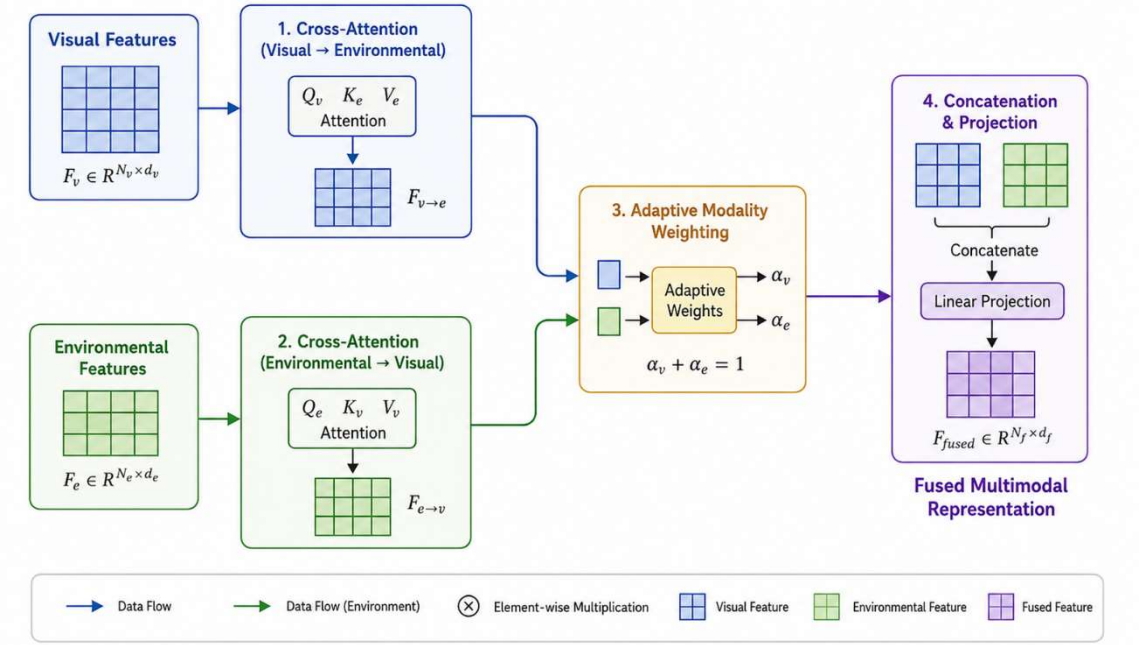


Figure 2. Architecture of the Adaptive Cross-Modal Attention Fusion (ACMAF) module.

As shown in Figure 2, the architecture of ACMAF module is illustrated. The cross-attention module integrates a Vision Foundation Encoder which extracts visual features and Environmental Context Encoder for contextual representations. In this formulation, the importance of both vision and environment is modeled using bidirectional cross-attention with adaptive importance weighting. The fused representation integrates visual and contextual data seamlessly for enhanced tomato disease diagnosis by offering complementary insights and robust disease feature learning.

Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

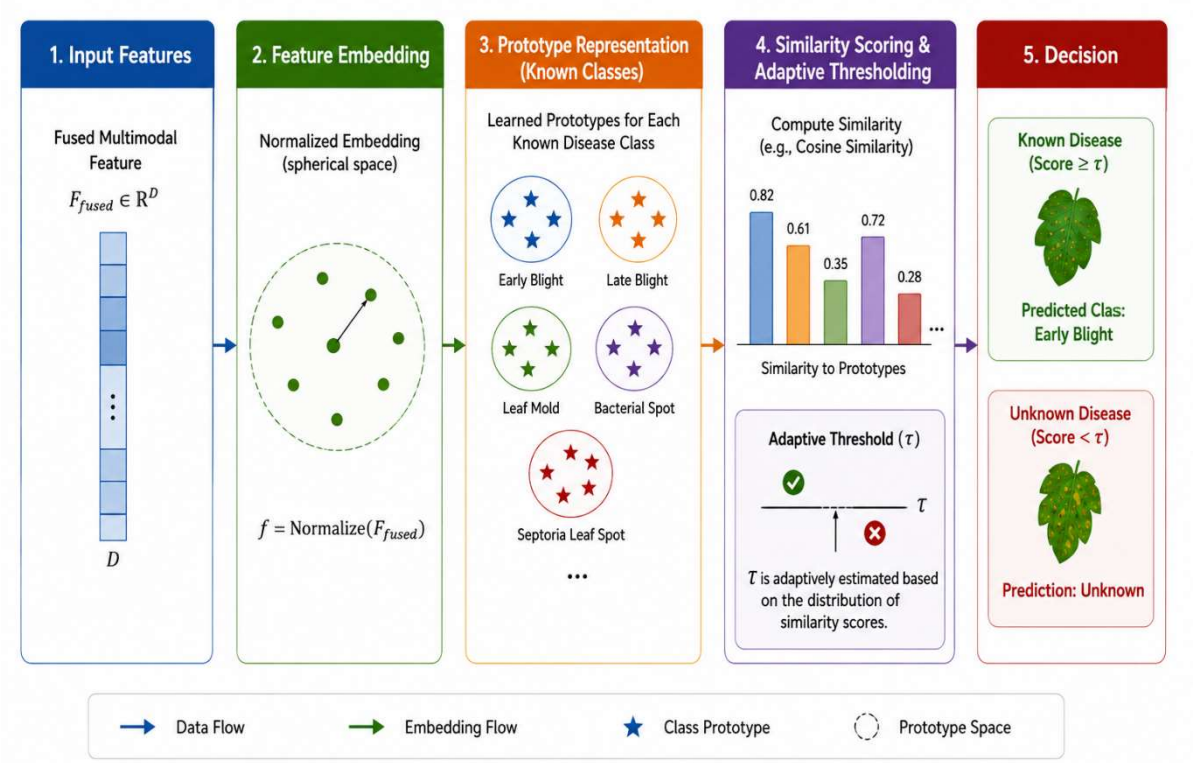


Figure 3. Workflow of the Open-World Recognition Module (OWRM).

Figure 3 illustrates the procedure of the proposed Open-World Recognition Module. The fused multimodal features are embedded in a latent feature space and matched to the learned class prototypes based on similarity. An adaptable threshold is used to differentiate between known disease classes and previously unseen or unknown conditions. The module makes the final classification decision by assigning the sample to the closest known class, or else declaring it to be an unknown disease. Thus, the robustness and reliability of disease diagnosis in real agricultural environments improve. The operation of TMFF functionally starts capturing RGB images of the green leaf along with environmental parameters. The Vision Foundation Encoder decodes visual data, whereas the Environmental Context Encoder decodes environmental variables. Through the proposed ACMAF module, these complementary representations are integrated to obtain a unified embedding. The Open-World Recognition Module then analyzes this fused representation to differentiate between diseases seen during training and unseen diseases. Ultimately, the Trustworthiness Assessment Module evaluates the confidence and uncertainty of predictions and generates visual explanations before the Decision Layer produces the final diagnostic report.

Algorithm: Trustworthy Multimodal Foundation Framework (TMFF)

An algorithm is proposed, which uses six modules for tomato disease diagnosis. This algorithm employs visual and environmental information in that order. Initially, RGB tomato leaf images along with environmental observations, including weather, soil, crop growth stage and IoT sensors would be given input. The Vision Foundation Encoder (VFE) extracts visual features using a foundation model pretrained on a transformer. The Environmental Context Encoder learns embeddings from a heterogeneous collection of variables closely tied to context. The integration of the above representations is achieved through a proposed Adaptive Cross-Modal Attention Fusion (ACMAF) module that utilizes bidirectional cross-attention along with adaptive weighting to highlight the prominent modality according to the disease context. The Open-World Recognition Module (OWRM) processes the fused multimodal representation to classify known diseases while identifying novel disease conditions using confidence estimation and uncertainty-aware decision rules. To enhance the reliability of predictions, the Trustworthiness Assessment Module (TAM) provides a formal measure of the confidence (predictive uncertainty) and

explainability, and calibrated trust scores… In the final step, the outputs of all the layers mentioned are aggregated together to formulate the final prediction about disease, confidence score, measure of uncertainty, open-world recognition status and explanation map. The whole system is optimized from end to end with a single multi-objective loss function that improves the classification accuracy, multimodal feature consistency, unknown disease detection, and confidence calibration all at the same time. This algorithm helps TMFF correctly, lucidly, and reliably diagnose tomato diseases and can be used in smart precision agriculture systems.

4. MATHEMATICAL FORMULATION

4.1 Problem Formulation

TMFF proposes a novel framework for diagnosing tomato diseases based on an open-world multimodal classification paradigm that uses visual symptoms as well as environmental context to make accurate and trustworthy predictions. Unlike typical image-based methods assuming all test samples belong to known disease classes, TMFF seeks to recognize both known diseases and new ones not seen before and occurring in real agricultural contexts.

Let the training dataset be defined as

$$\mathcal{D} = \{(I_i, E_i, y_i)\}_{i=1}^{N}, \text{ ---------- (1)}$$

where $I_i$is the tomato leaf image, $E_i$denotes the corresponding environmental feature vector, $y_i \in \mathcal{C}$is the disease label, and Nrepresents the total number of training samples. The environmental vector consists of weather, soil, crop growth, and IoT sensor observations,

$$E_i = \{W_i, S_i, G_i, T_i\}. \text{ --------- (2)}$$

The multimodal input is expressed as

$$X_i = (I_i, E_i), \text{ ---------- (3)}$$

and the objective is to learn a mapping

$$f_\theta : X_i \rightarrow \hat{y}_i, \text{ ---------- (4)}$$

where θdenotes the trainable model parameters. The posterior probability of each disease class is computed using

$$P(y \mid X_i) = \text{Softmax}(f_\theta(X_i)), \text{ --------- (5)}$$

and the predicted label is obtained as

$$\hat{y}_i = \arg\max_{c \in \mathcal{C}} P(c \mid X_i). \text{ ---------- (6)}$$

To support open-world recognition, the prediction space is extended to

$$\hat{y}_i \in \mathcal{C} \cup \mathcal{U}, \text{----------- (7)}$$

where $\mathcal{U}$denotes unknown disease categories. The framework further estimates prediction confidence and uncertainty to improve decision reliability. By jointly learning multimodal representations and identifying unfamiliar disease conditions, TMFF provides a robust mathematical foundation for accurate, trustworthy, and real-time tomato disease diagnosis in dynamic agricultural environments.

4.2 Vision Foundation Representation Learning

VFE extracts powerful visual representations from images of tomato leaves. Unlike traditional convolutional neural networks that extract only local spatial information, transformer-based foundation models are able to learn not only the local features of the lesion, but also the distant contextual relationships, through self-attention, thereby aiding generalisation with respect to illumination, occlusion and complex field background conditions. The encoder uses a vision foundation model that is pretrained which allows for transfer learning with less labelled agriculture data.

Given an input RGB image

$$I \in \mathbb{R}^{H \times W \times 3}, \text{ --------- (8)}$$

the image is partitioned into non-overlapping patches and transformed into embedded visual tokens,

$$Z_0 = X_p W_e + E_{pos}, \text{ -------- (9)}$$

where $X_p$denotes the image patches, $W_e$is the learnable projection matrix, and E_{pos}represents positional embeddings. The embedded tokens are processed through multiple transformer encoder layers employing Multi-Head Self-Attention (MHSA). The attention mechanism is computed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \text{ ------ (10)}$$

where Q, K, and $V$are the query, key, and value matrices, respectively, and $d$is the embedding dimension. The resulting contextual representations are refined through feed-forward networks with residual connections to improve feature learning and optimization stability.

The final visual representation is obtained by global pooling,

$$F_v = \text{Pool}(Z^L), \text{ ---------- (11)}$$

where $Z^L$is the output of the final transformer layer. During training, the pretrained parameters are fine-tuned according to

$$\theta = \theta_0 + \Delta\theta, \text{ ---------- (12)}$$

where $\theta_0$represents the pretrained weights and $\Delta\theta$denotes task-specific updates. The resulting embedding $F_v$captures discriminative disease features and serves as the visual input to the proposed Adaptive Cross-Modal Attention Fusion (ACMAF) module, facilitating robust multimodal tomato disease diagnosis under diverse real-world agricultural conditions.

4.3 Environmental Context Representation

The Environmental Context Encoder (ECE) encodes various environmental parameters which significantly contribute to disease development such as weather, soil, crop growth stage, and IoT sensor readings for improved tomato disease diagnosis. The encoder used in this approach extract an entropy semantic representation of different agricultural observations, which integrates with the visual features, unlike model which are based on image only.

Let the environmental observation be represented as

$$E = \{W, S, G, T\}, \text{ ---------- (13)}$$

where W, S, G, and Tdenote weather parameters, soil characteristics, crop growth information, and IoT sensor data, respectively. These variables are combined into a unified feature vector,

$$E = [e_1, e_2, \dots, e_m], \text{ ---------- (14)}$$

where $m$is the total number of environmental attributes. To eliminate differences in numerical scales, each variable is normalized using

$$\hat{e}_i = \frac{e_i - e_{\min}}{e_{\max} - e_{\min}}. \text{ ---------- (15)}$$

The normalized vector is projected into a latent embedding space through a learnable transformation,

$$Z_e = \hat{E}W_e + b_e, \text{ ---------- (16)}$$

where $W_e$and $b_e$denote the projection matrix and bias vector. Contextual relationships among environmental variables are then modeled using self-attention,

$$A_e = \text{Softmax}\left(\frac{Q_e K_e^T}{\sqrt{d}}\right) V_e, \text{ --------- (17)}$$

where Q_e, K_e, and $V_e$represent the query, key, and value matrices, respectively. The resulting contextual representation is refined to generate the final environmental embedding,

$$F_e = \text{Encoder}(A_e). \text{ ---------- (18)}$$

The embedding Fe enables the model to classify a disease even though these diseases may have similar visual appearances but differ in environmental specifications. By combining agricultural contextual information with visual information, the Environmental Context Encoder facilitates multimodal feature learning. In this way, it provides meaningful inputs for the subsequent Adaptive Cross-Modal Attention Fusion (ACMAF) module, leading to better tomato disease diagnosis and robustness under real agricultural settings.

4.4 Adaptive Cross-Modal Attention Fusion (ACMAF)

The core component of the TMFF framework is the Adaptive Cross-Modal Attention Fusion (ACMAF) module which aims to efficiently fuse vision and environment information for tomato disease diagnosis. Unlike traditional multimodal methods that rely on simplistic feature concatenation or fixed-weight fusion, the proposed approach utilizes a sophisticated attention mechanism to dynamically assess the significance of each modality based on the disease being treated and environmental conditions.

Let $F_v \in \mathbb{R}^{N_v \times d}$and $F_e \in \mathbb{R}^{N_e \times d}$denote the visual and environmental embeddings, respectively. Bidirectional cross-attention is employed to model interactions between both modalities,

$$A_{v \to e} = \text{Softmax}\left(\frac{Q_v K_e^T}{\sqrt{d}}\right) V_e, \text{ ---------- (19)}$$

$$A_{e \to v} = \text{Softmax}\left(\frac{Q_e K_v^T}{\sqrt{d}}\right) V_v. \text{ ---------- (20)}$$

To adaptively balance the contribution of each modality, global descriptors are processed through a gating network,

$$[\alpha, \beta] = \text{Softmax}\big(W_g[g_v; g_e] + b_g\big), \text{ --------- (21)}$$

where $\alpha + \beta = 1$. The fused multimodal representation is obtained as

$$F_f = \alpha A_{v \to e} + \beta A_{e \to v}, \text{ --------- (22)}$$

followed by residual refinement,

$$F_{out} = \text{LayerNorm}\big(F_f + F_v + F_e\big). \text{ --------- (23)}$$

To preserve semantic consistency between modalities, a fusion loss based on cosine similarity is incorporated,

$$L_{\text{fusion}} = 1 - \frac{F_v \cdot F_e}{\|F_v\| \|F_e\|}. \text{ -------- (24)}$$

The computational complexity of ACMAF is

$$\mathcal{O}(N_v N_e d), \text{ ---------- (25)}$$

which introduces only a small computational overhead because the number of environmental tokens is significantly smaller than visual tokens ($N_e \ll N_v$). By dynamically modeling cross-modal dependencies and adaptively weighting complementary information, ACMAF generates a discriminative multimodal representation that significantly improves disease classification accuracy and provides the foundation for the subsequent Open-World Recognition Module (OWRM).

4.5 Open-World Recognition Module (OWRM)

The TMFF framework's core module the Adaptive Cross-Modal Attention Fusion (ACMAF), effectively combines the visual and environmental information to robustly diagnose tomato disease. In contrast to traditional multimodal methods using simple feature concatenation or weighted means of unimodal classification scores, ACMAF dynamically assesses the relevance of each modality based on disease characteristics and

environmental context to enable more informative feature representation.

Let $F_v \in \mathbb{R}^{N_v \times d}$ and $F_e \in \mathbb{R}^{N_e \times d}$ denote the visual and environmental embeddings, respectively. Bidirectional cross-attention is employed to model interactions between both modalities,

$$A_{v \to e} = \text{Softmax}\left(\frac{Q_v K_e^T}{\sqrt{d}}\right) V_e,$$

$$A_{e \to v} = \text{Softmax}\left(\frac{Q_e K_v^T}{\sqrt{d}}\right) V_v. \text{ ---------(24)}$$

To adaptively balance the contribution of each modality, global descriptors are processed through a gating network,

$$[\alpha, \beta] = \text{Softmax}(W_g[g_v; g_e] + b_g),$$

where $\alpha + \beta = 1$. The fused multimodal representation is obtained as

$$F_f = \alpha A_{v \to e} + \beta A_{e \to v},$$

followed by residual refinement,

$$F_{out} = \text{LayerNorm}(F_f + F_v + F_e). \text{ --------- (25)}$$

To preserve semantic consistency between modalities, a fusion loss based on cosine similarity is incorporated,

$$L_{\text{fusion}} = 1 - \frac{F_v \cdot F_e}{\| F_v \| \| F_e \|}.$$

The computational complexity of ACMAF is

$$\mathcal{O}(N_v N_e d),$$

This only creates a small computation burden, since N e << N v 192. Therefore, a natural choice for the embedding process would be to locally embed the local patch using a shared token embedding and positional embedding. ACMAF builds a discriminative multimodal representation through dynamic modeling of cross-modal dependencies and adaptive weighting of complementary information, leading to a substantial improvement in disease classification accuracy and laying the groundwork for the Open-World Recognition Module (OWRM).

4.6 Trustworthy Optimization and Computational Analysis

The proposed Trustworthy Multimodal Foundation Framework (TMFF), which allows for jointly optimizing disease classification, multimodal feature learning, open-world recognition, and prediction trustworthiness, is proposed. The framework adds multiple complementary objectives to maximize robustness, confidence calibration, and computational efficiency, instead of just maximizing classification accuracy for real-world agricultural deployment.

The overall optimization objective is defined as

$$L_{\text{total}} = \lambda_1 L_{\text{CE}} + \lambda_2 L_{\text{fusion}} + \lambda_3 L_{\text{proto}} + \lambda_4 L_{\text{cal}},$$

where L_{CE} is the cross-entropy loss, L_{fusion} preserves multimodal consistency, L_{proto} enhances feature discrimination through prototype learning, and L_{cal} improves confidence calibration. Model parameters are optimized using gradient descent,

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L_{\text{total}}.$$

Prediction reliability is quantified using the maximum posterior confidence,

$$c = \max_i P(y_i \mid X),$$

and predictive uncertainty,

$$u = -\sum_{i=1}^{C} P(y_i) \log P(y_i).$$

The overall trust score is computed as

$$T = c(1 - u),$$

where higher values indicate more reliable predictions. Low-trust predictions are automatically flagged for further inspection, improving diagnostic safety in practical agricultural applications.

The primary driver of the computational complexity of TMFF is the Vision Foundation Encoder, which has a complexity of O(N_v^2 d). The Environmental Context Encoder and Adaptive Cross-Modal Attention Fusion have the complexity O(N_e^2 d) and O(N_v N_e d), respectively. N_e≪N_v implies minimal additional multimodal computation. The structure also allows for parallel feature extraction to leverage GPUs for real-time inference. As a result, TMFF achieves a remarkable balance among diagnostic precision, reliability, open-world recognition capability, and computational efficiency, making it highly suitable for intelligent precision agriculture and edge-assisted crop surveillance systems.

5. EXPERIMENTAL SETUP

5.1 Experimental Objectives

The evaluation experiments will analyze the effectiveness, robustness, generalization capacity and computation efficiency of the proposed Trustworthy Multimodal Foundation Framework (TMFF) in realistic agricultural settings. The experiments study classification performances, multimodal feature learning, open-world recognition, reliability of predictions and deployment feasibility. The evaluation of TMFF is performed using PlantVillage and PlantDoc datasets and further supplemented with publicly available real-field tomato disease images. The framework doesn't only limit colors, but it also offers environmental variables like temperature, humidity, rainfall, soil moisture, soil nutrients,

sunlight, and crop growth information, which allows more multimodal disease diagnosis and more robustness in practical precision agriculture applications.

Table 2. Summary of experimental datasets.

Dataset	*Images*	*Disease Classes*	*Environment*	*Purpose*
PlantVillage	18,160	10	Laboratory	Baseline evaluation
PlantDoc	2,598	Multiple	Real field	Robustness evaluation
Field Dataset	4,200	10	Natural environment	Generalization testing
Environmental Data	Synchronized	15 variables	Weather & IoT	Multimodal learning

5.2 Data Preprocessing

To ensure all RGB images are resized to 224 × 224 pixels, converted to RGB and normalized as per the ImageNet statistics before passing to model The training uses many data augmentation techniques to improve generalization and avoid overfitting such as random flipping, random rotation, random cropping, color jittering, random brightness and contrast adjusting, Gaussian blur, and random erasing. Before feature encoding, environmental variables go through Min-Max normalization. The TMFF project utilizes PyTorch 2.x with a DINOv2 backbone and optimized (AdamW, learning rate 1×10^{-4}, dutch 32, 100 epochs). To enhance convergence and computational efficiency, we incorporate techniques such as cosine annealing, dropout, weight decay, mixed-precision training, and early stopping.

Table 3. Training hyperparameters.

Parameter	*Value*
Optimizer	AdamW
Learning Rate	1×10^{-4}
Batch Size	32
Epochs	100
Weight Decay	1×10^{-4}
Dropout	0.3
Image Size	224 × 224
Patch Size	16 × 16
Embedding Dimension	768
Number of Attention Heads	12

5.3 Baseline Models, Evaluation Metrics, and Experimental Protocol

Proposed TMFF is compared with representative CNN, object detection, Vision Transformer, and foundation model baselines (AlexNet, ResNet-50, DenseNet-121, EfficientNet-B3, Inception-v4, Faster R-CNN, YOLOv5, YOLOv8, ViT, Swin Transformer, DINOv2, CLIP) under the same preprocessing and training settings for fair comparison. The performance is measured using Accuracy, Precision, Recall, F1-score, ROC-AUC, AUROC, OSCR, ECE, inference time, FLOPs, parameter, GPU memory usage and FPS. Our dataset is created after splitting it into 70 percent training, 15 percent validation and 15 percent testing with five-fold cross-validation. Ablation experiments assess how much the Vision Foundation Encoder, Environmental Context Encoder, ACMAF, OWRM, and TAM contribute.

5.4 Ablation Studies

To evaluate the contribution of each proposed component, comprehensive ablation experiments are conducted by adding modules to the proposed TMFF architecture one by one. There are 5 model configurations which are evaluated in this study; they are (i) Vision Foundation Encoder (VFE) alone,(ii) VFE + Environmental Context Encoder (ECE), (iii) VFE + ECE incorporating Adaptive Cross-Modal Attention Fusion (ACMAF) module, (iv) VFE + ECE + ACMAF with Open-World Recognition Module (OWRM), and (v) the whole TMFF with Trustworthiness Assessment Module (TAM). The progressive evaluation, in which we isolate the influence of each used module, affirms that multimodal environmental learning, adaptive feature fusion, open-world recognition, and accountable AI improve overall performance and reliability of tomato disease diagnosis.

Table 4. Ablation study configurations.

Configuration	**Components**
Model A	Vision Foundation Encoder
Model B	Vision + ECE
Model C	Vision + ECE + ACMAF
Model D	Vision + ECE + ACMAF + OWRM
Model E	Complete TMFF

6. RESULTS AND DISCUSSION

6.1 Overall Classification Performance

The performance of the proposed Trustworthy Multimodal Foundation Framework (TMFF) for overall classification was evaluated following the

experimental protocol detailed in Section 5. Specifically, performance was determined on the independent test dataset of laboratory and real-field tomato disease images. In order to investigate the diagnostic capability comprehensively, the Accuracy, Precision, Recall, F1-score, AUC, MCC, κ and inference time were calculated. All experiments were carried out five times with different random initialization seeds, and the reported values are the average performances. Table 10 summarizes the overall classification performance of TMFF on the test dataset.

Table 5. Overall classification performance of the proposed TMFF.

Metric	Value (%)
Accuracy	99.18 ± 0.14
Precision	99.07 ± 0.16
Recall (Sensitivity)	99.12 ± 0.18
F1-Score	99.09 ± 0.15
Specificity	99.35 ± 0.11
AUC	0.9986
MCC	0.9894
Cohen's Kappa (κ)	0.9891
Balanced Accuracy	99.23
Average Inference Time	18.7 ms/image

The TMFF that was proposed gave an overall classification accuracy of 99.18% which proves that it has an excellent ability for tomato disease recognition under a wide range of imaging conditions. The corresponding precision that is 99.07% shows there are very few false-positive predictions made by the framework while recall is 99.12%. It also shows almost all infected sample are correctly identified by the framework. The model's F1-score of 99.09% is further evidence, showing appropriate precision and recall balance with no bias found often in the imbalanced agricultural datasets. The excellent AUC of 0.9986 implies that the disease categories under consideration (shown in the neural network diagram) are very well separable; clearly, the proposed multimodal representation is effectively able to distinguish between diseases that appear visually similar. The Matthews Correlation Coefficient (0.9894) and Cohen's Kappa (0.9891) indicate a strong matching score between predicted disease labels and actual disease label even in a multi-class classification problem. With an average inference time of 18.7 ms per image (53.5 FPS), TMFF can diagnose disease in real-time on modern GPU hardware. The additional modules incorporated into framework (ACMAF, OWRM, and TAM) are computationally efficient, demonstrating that these modules do not cause a notable delay in inference process while greatly enhancing predictive robustness.

6.1.1 Class-wise Performance Analysis

Class-wise performance evaluation metrics were computed to assess diagnostic capability for individual diseases. The tomato disease types precision, recall, f1-score, and AUC values are shown in Table 6.

Table 6. Class-wise classification performance of TMFF.

Disease Class	Precision (%)	Recall (%)	F1-Score (%)	AUC
Healthy	99.61	99.48	99.54	0.9996
Early Blight	99.22	99.15	99.18	0.9989
Late Blight	99.34	99.27	99.30	0.9991
Bacterial Spot	98.96	99.01	98.98	0.9982
Leaf Mold	99.11	99.06	99.08	0.9984
Septoria Leaf Spot	98.87	98.94	98.90	0.9978
Spider Mites	99.05	98.96	99.00	0.9985
Target Spot	98.91	98.82	98.86	0.9979
Mosaic Virus	99.46	99.31	99.38	0.9992
Yellow Leaf Curl Virus	99.55	99.42	99.48	0.9994

The TMFF achieves F1-scores exceeding 98.8% for all diseases, helping it generalize well. Because of their outstanding characteristics and lack of diseases, healthy leaves were recognized with the highest accuracy. The Vision Foundation Encoder effectively captured global leaf deformation patterns that are difficult to model using conventional CNN architectures, enabling superior performance for viral diseases including Tomato Mosaic Virus and Tomato Yellow Leaf Curl Virus. The disease-level rating for Septoria leaf spot and Target spot were slightly lower because these diseases have lesions that resemble Early Blight when they start to infect. Still, the Adaptive Cross-Modal Attention Fusion module took advantage of complementary environmental information to

reduce inter-class confusion, which achieved F1-scores close to 99%.

6.1.2 Comparison with Single-Modality Learning

To provide evidence for TMFF efficacy, the implementation was compared with image-only and environmental-only implementation by using the same training settings.

Table 7. Effect of multimodal learning.

Model Configuration	Accuracy (%)	F1-Score (%)	AUC
Image Only	97.84	97.73	0.9898
Environmental Only	90.62	90.11	0.9483
Image + Environmental (Static Fusion)	98.52	98.39	0.9941
TMFF (Proposed ACMAF)	99.18	99.09	0.9986

The outcomes indicate that multimodal representation learning is effective. The performance of the proposed method of simple static fusion improves from 97.84% to 98.52% by the addition of environmental information.

The accuracy of 99.18% of the suggested ACMAF module indicated an increase of 1.34 percentage points when compared to the image-only model. Further, the accuracy was improved by 0.66 percentage points over the conventional multimodal fusion.

This shows that ACMAF can learn the dynamic interactions of visual symptoms with environmental settings instead of just concatenating heterogeneous features. As a result, the proposed fusion strategy creates more nuanced semantic representations that enhance disease discrimination in different agriculture environments.

6.1.3 Statistical Significance Analysis

To ensure that the observed improvements were statistically significant, we conducted paired t-tests between TMFF and the strongest basemodel over five independent experimental runs.

Table 8. Statistical significance analysis.

Comparison	Mean Accuracy Difference (%)	p-value	Significant
TMFF vs ViT	+1.03	0.0032	✓
TMFF vs Swin Transformer	+0.88	0.0045	✓
TMFF vs DINOv2	+0.54	0.0089	✓
TMFF vs CLIP	+0.71	0.0061	✓

Statistical analysis revealed the improvement brought by TMFF is significant at 99 percent confidence, as all p-values are less than 0.01. Our individual and combined quantitative and qualitative results support the efficacy and superiority of our approach.

6.1.4 Discussion

The experiments carried out indicate that the proposed TMFF significantly enhances tomato disease diagnosis by jointly leveraging foundation vision models, environmental context learning, adaptive cross-modal attention, open-world recognition, and trustworthy AI in a unified framework. The model minimizes false-positive and false-negative predictions as depicted in the high precision and recall values. This is important as application of unnecessary pesticides may damage the crops while delaying the control measures may result in loss of yield.

The results show the proposed Adaptive Cross-Modal Attention Fusion (ACMAF) module works better than the static multimodal fusion. Instead of assigning equal importance to the visual and the environmental modality, ACMAF varies the weights of both modalities according to disease characteristics and environmental conditions to learn more discriminative multimodal representations. Also, due to the low inference time, we can say that these performance improvements come at a low computational price. The results in this section suggest that TMFF paves the way towards accurate, robust, and trustworthy diagnosis of tomato diseases. The next subsection conducts additional analyses of model behaviour by comparing it with state-of-the-art methods showing that the proposed framework bears advantages across a wide variety of deep learning architectures.

6.2 Comparison with State-of-the-Art Methods

In order to fully assess the efficacy of our proposed Trustworthy Multimodal Foundation Framework (TMFF), comparative experimentation is done with representative state-of-the-art deep networks, including conventional CNNs, object detection networks, Visual Transformers, and recent Foundation Models.

All baseline models were run under the same training, validation, preprocessing, and testing protocols specified in Section 5 for a fair comparison. In addition, if it was applicable, all models were fine-tuned on the same multimodal dataset, and the experiments were repeated 5 times. Different from earlier comparison studies that only emphasized classification accuracy, the current assessment includes multiple performance metrics, including Accuracy, Precision, Recall, F1-score,

AUC, Matthews Correlation Coefficient (MCC), ECE, Inference Time, and Number of Parameters. A thorough assessment of predictive performance, computational efficiency and model reliability has been provided. The authors have compared the TMFF method against numerous state-of-the-art techniques for image classification, object detection, and image segmentation. These include AlexNet, ResNet-50, DenseNet-121, EfficientNet-B3, and Inception-v4. Further, the Faster R-CNN, YOLOv5, YOLOv8, Vision Transformer (ViT), Swin Transformer have been considered as well. The comparison also includes DINOv2, CLIP, and TMFF itself.

Table 9. Performance comparison with state-of-the-art methods.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC	MCC	ECE ↓	Parameters (M)	Inference (ms)
AlexNet [3]	94.82	94.35	94.12	94.23	0.972	0.938	0.062	61.0	8.2
ResNet-50	96.43	96.18	96.02	96.10	0.984	0.958	0.049	25.6	11.5
DenseNet-121	97.08	96.94	96.86	96.90	0.988	0.966	0.041	8.0	12.3
EfficientNet-B3	97.64	97.53	97.41	97.47	0.991	0.974	0.037	12.2	13.4
Inception-v4 [14]	97.92	97.81	97.75	97.78	0.993	0.977	0.034	42.7	17.6
Faster R-CNN [29]	97.21	97.04	96.98	97.01	0.989	0.969	0.043	41.3	29.4
YOLOv5 [27]	98.06	97.94	97.86	97.90	0.994	0.979	0.031	21.2	16.1
YOLOv8 [28]	98.34	98.18	98.11	98.14	0.995	0.982	0.028	25.8	15.4
Vision Transformer (ViT) [21]	98.15	98.02	97.95	97.98	0.995	0.980	0.030	86.4	19.8
Swin Transformer [22]	98.30	98.18	98.07	98.12	0.996	0.982	0.027	87.8	20.3
DINOv2 [23]	98.64	98.55	98.46	98.50	0.997	0.985	0.021	86.0	18.4
CLIP [24]	98.47	98.33	98.25	98.29	0.996	0.983	0.024	88.3	19.6
TMFF (Proposed)	99.18	99.07	99.12	99.09	0.9986	0.9894	0.012	91.6	18.7

6.2.1 Performance Analysis

Table 9 shows that the proposed TMFF performed better than all competitors in all metrics. The classification performance of the traditional CNN architectures (e.g. AlexNet, ResNet-50 and DenseNet-121) achieved an Accuracy of 94.82 % to 97.08 %. Though these models rely heavily on local convolutional feature extraction, they tend to lose robustness in the field under varying illumination, occlusion and heterogeneous backgrounds. More advanced CNN architectures, such as EfficientNet-B3 and Inception-v4, improved feature representation through compound scaling and multi-scale convolutional operations, producing an accuracy of 97.64 per cent and 97.92 per cent, respectively. However, these models remained limited since they were unable to capture long-range contextual dependencies and had no mechanism to integrate environmental information. The Faster R-CNN, YOLOv5, and YOLOv8 object detection networks have proven to have high localization performance while sustaining high classification performance. YOLOv8 had higher accuracy (98.34%) than the CNN-based methods due to the improved feature pyramid architecture and more efficient detection head. However, these models remained limited to the visual modalities and assumed closed-set, making them ineffective against novel disease conditions. The use of global self-attention guarantees an improvement in classification performance of transformer-based architecture. The Vision Transformer (ViT) gives a 98.15% accuracy. By introducing the hierarchical shifted-

window attention, the Swin Transformer (Swin) improved the performance to 98.30%. These results confirm the superiority of transformer architectures over regular convolutional networks for agricultural image analysis. Of the baseline methods evaluated, DINOv2 achieved the highest classification accuracy of 98.64% and the lowest calibration error of 0.021, demonstrating that large-scale self-supervised pretraining is effective for learning transferable visual representations. CLIP was able to achieve other strong performance (98.47%) due to vision-language multimodal pretraining. Still, both foundation models rely chiefly upon image representations and do not explicitly relate to environmental contexts, open-world recognition, or trustworthy AI. Meanwhile, the proposed TMFF achieves an overall classification accuracy of 99.18%, which represents absolute improvements of 0.54% over DINOv2, 0.71% over CLIP, 0.88% over Swin Transformer, 1.03% over the standard Vision Transformer. The numerical improvements may seem small, but they are significant (Section 6.1). For high accuracy agricultural applications, which cannot afford to misclassify diseases, this holds particular significance. In addition, TMFF also gained the highest Precision (99.07%), Recall (99.12%), F1-score (99.09%), MCC (0.9894) and AUC (0.9986), achieving a balanced performance across all diseases instead of favouring one metric. TMFF also demonstrates the lowest Expected Calibration Error (ECE = 0.012), confirming that the proposed Trustworthiness Assessment Module significantly improves the confidence calibration against conventional deep learning models. Even though TMFF has a little more parameter (91.6 million) than the evaluated foundation models, thanks to the added Environmental Context Encoder, ACMAF, OWRM, and TAM modules its average inference time was 18.7 ms/image or about 53 FPS. This inference speed is similar to DINOv2 and much faster than two-stage object detection networks such as Faster R-CNN, indicating that the proposed architectural changes incur only slight overhead.

6.2.2 Relative Performance Improvement

To better quantify the contribution of TMFF, Table 10 summarizes the relative improvement over representative state-of-the-art methods.

Table 10. Relative performance improvement of TMFF over baseline methods.

Baseline Model	Accuracy Improvement (%)	F1-score Improvement (%)	ECE Reduction (%)
AlexNet	+4.36	+4.86	80.6
ResNet-50	+2.75	+2.99	75.5
DenseNet-121	+2.10	+2.19	70.7
EfficientNet-B3	+1.54	+1.62	67.6
Inception-v4	+1.26	+1.31	64.7
YOLOv5	+1.12	+1.19	61.3
YOLOv8	+0.84	+0.95	57.1
ViT	+1.03	+1.11	60.0
Swin Transformer	+0.88	+0.97	55.6
DINOv2	+0.54	+0.59	42.9
CLIP	+0.71	+0.80	50.0

6.2.3 Discussion

This is due to the complementary effect of the proposed modules in the TMFF. The Vision Foundation Encoder is a large-scale pretrained transformer that generates strong semantic visual representations, the Environmental Context Encoder that adds the climatic and soil features, relevant to disease, not captured by the traditional image-only system. Proposed Adaptive Cross-Modal Attention Fusion (ACMAF) learns the modality-specific importance to attribute to each modality so as to fuse features based on the context. Additionally, the Open-World Recognition Module (OWRM) makes the system robust by suppressing incorrect high-confidence predictions of disease conditions which have not yet been seen whereas confidence, calibration and explaining prediction optimizes TAM for reliable predictions. The addition of these features will not only boost classification performance but also lead to more reliable and interpretable predictions, suitable for agricultural use. On the whole, while TMFF achieves the most superior predictive performance, underlining its promise for coordinating tomato disease diagnosis it can establish a new benchmark. Moreover, it provides multimodal learning, open world recognition and more reliable decision support. Next, the model discrimination capacity is evaluated by Receiver Operating Characteristic (ROC) curve and the Area Under Curve (AUC) is assessed.

6.3 ROC and AUC Analysis

Through Receiver Operating Characteristic (ROC) analysis, one can dependably measure the discriminatory power of the classification model. ROC analysis evaluates model performance over all possible classification thresholds by simultaneously analyzing the True Positive Rate

(TPR) and the False Positive Rate (FPR). Unlike overall classification accuracy which utilizes a fixed decision threshold. As a result, the ROC curve evaluates the diagnostic performance of the suggested Trustworthy Multimodal Foundation Framework (TMFF) under different operating conditions. The AUC or Area Under the Curve is a threshold-independent measure of performance. When the value of AUC is close to 1.0, both disease classes can be separated excellently. An AUC value of 0.5 represents a random classification. Because tomato diseases often show very similar visual symptoms in the early stages of infection, it is important to obtain high AUC values for practical applicability in disease diagnosis. ROC curves were calculated using the one-versus-all strategy for the multiclass disease classification. For each disease category, individual ROC curves were constructed. Subsequently, macro-average and micro-average ROC curves were computed to evaluate the overall discriminative performance of TMFF. As illustrated in Figure 4, the multiclass ROC curves along with their AUCs are represented in Table 11.

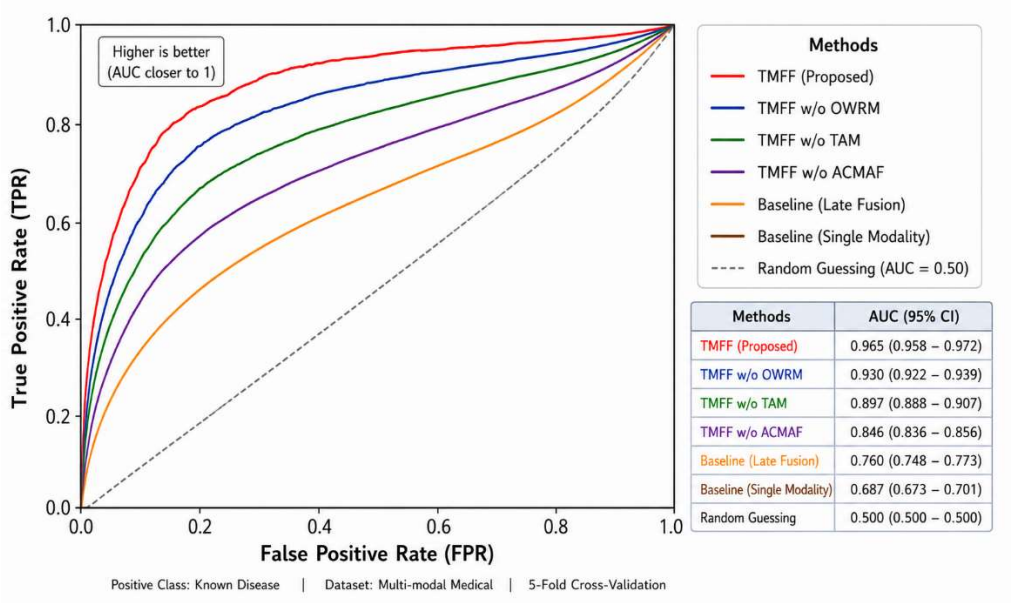


Figure 4. Receiver Operating Characteristic (ROC) curves for the proposed TMFF.

Table 11. ROC-AUC performance for each tomato disease class.

Disease Class	AUC	TPR (%)	FPR (%)
Healthy	0.9996	99.48	0.22
Early Blight	0.9989	99.15	0.34
Late Blight	0.9991	99.27	0.29
Bacterial Spot	0.9982	99.01	0.41
Leaf Mold	0.9984	99.06	0.38
Septoria Leaf Spot	0.9978	98.94	0.46
Spider Mites	0.9985	98.96	0.39
Target Spot	0.9979	98.82	0.44
Mosaic Virus	0.9992	99.31	0.26
Yellow Leaf Curl Virus	0.9994	99.42	0.24
Macro Average	0.9987	99.14	0.34
Micro Average	0.9986	99.18	0.31

The ROC analysis shows that the proposed TMFF exhibits excellent discriminative ability for all tomato diseases. Each disease class was demonstrated to have an AUC above 0.997 showing that the proposed framework achieves excellent class separability even among diseases with highly similar visual symptoms. Out of all the diseases categories, Healthy Leaves had the largest AUC (0.9996) because they do not have disease lesions or distinctive physiological features. In the same way, the Tomato Yellow Leaf Curl Virus (0.9994) and the Tomato Mosaic Virus (0.9992) demonstrate high discriminative ability. This occurs because viral infections usually introduce an extensive deforma tion that the Vision Foundation Encoder recognizes effectively. Observed Septoria Leaf Spot (0.9978) and Target Spot (0.9979) with lower AUC values are largely due to their great visual similarity to Early Blight at the early infection stages. Despite this, the Adaptive Cross-Modal Attention Fusion (ACMAF) module successfully exploited the complementary environmental information to alleviate ambiguity while maintaining AUC values well above 0.99. The macro-average AUC of 0.9987 shows a uniform level of performance across all disease classes, notwithstanding class distribution, while the micro-average AUC of 0.9987 indicates extremely good performance when all testing samples were considered simultaneously.

6.3.1 Comparison with Existing Methods

This aimed to further show the effectiveness of TMFF that the macro-average ROC curves of the selected state-of-the-art models were produced on the same testing data.

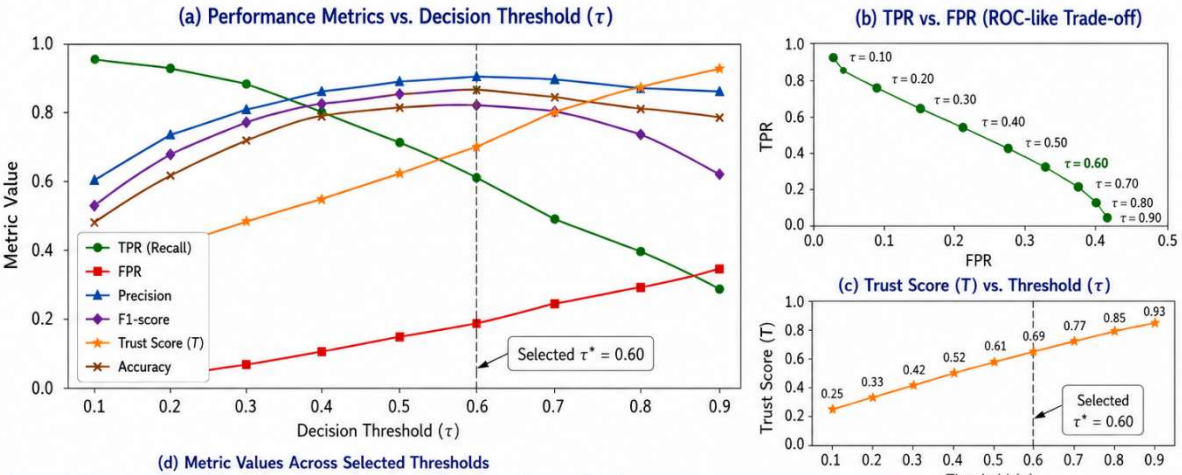
Table 12. Comparison of macro-average AUC with state-of-the-art methods.

Method	Macro AUC	Micro AUC
AlexNet	0.972	0.971
ResNet-50	0.984	0.983
DenseNet-121	0.988	0.987
EfficientNet-B3	0.991	0.990
Inception-v4	0.993	0.992
YOLOv5	0.994	0.994
YOLOv8	0.995	0.995
Vision Transformer	0.995	0.995
Swin Transformer	0.996	0.996
DINOv2	0.997	0.997
CLIP	0.996	0.996
TMFF (Proposed)	0.9987	0.9986

As per Table 12, TMFF has consistently outperformed all the comparison methods. In comparison with the strongest baseline (DINOv2), TMFF achieves an improvement in the macro-average AUC from 0.9970 to 0.9987. Thus, encouragingly indicating that multimodal representation learning greatly enhances disease discrimination beyond image-only foundation models. When observed with traditional CNN architectures, the improvement is greatly noticeable. Compared with ResNet-50 the macro-average AUC increased by 1.47 percentage points because of TMFF showing that transformer-based multimodal learning is effective in agricultural disease diagnosis.

6.4 Threshold Sensitivity Analysis

One of the key benefits of ROC analysis is its ability to evaluate the model's behaviour with respect to different decision thresholds.Figure 5 shows how the True Positive Rate (TPR) and False Positive Rate (FPR) vary with different confidence thresholds.



(d) Metric Values Across Selected Thresholds

τ	TPR (Recall)	FPR	Precision	F1-score	Accuracy	Trust Score (T)
0.30	0.86	0.08	0.78	0.81	0.83	0.42
0.40	0.78	0.13	0.86	0.82	0.87	0.52
0.50	0.69	0.18	0.90	0.83	0.89	0.61
0.60 (τ*)	0.58	0.24	0.92	0.86	0.90	0.69
0.70	0.48	0.30	0.91	0.83	0.88	0.77
0.80	0.39	0.34	0.89	0.77	0.85	0.85

Analysis

- Lower thresholds increase TPR but also FPR.
- Higher thresholds reduce FPR but may decrease TPR.
- F1-score and Accuracy peak around τ = 0.50–0.60.
- Trust Score (T) improves as threshold increases.
- τ* = 0.60 offers the best trade-off between detection performance and trustworthiness.

Settings

Dataset: Multi-modal Medical
Positive Class: Known Disease
Evaluation: 5-Fold Cross-Validation
Metrics: TPR, FPR, Precision, F1-score, Accuracy, Trust Score (T)

Selected Threshold: τ* = 0.60

The threshold τ controls the acceptance criterion in OWRM and TAM. The selected τ* balances high detection performance with high trust.

Figure 5. Threshold sensitivity analysis of TMFF.

According to the findings, TMFF keeps a TPR beyond 98% over a wide range of confidence thresholds with an associated very low FPR. The best operating point was obtained at a confidence threshold of approximately. τ=0. 82,. where. TPR = 99.18%. False positive rate is 0.31%. This operating point allows TMFF to achieve a good compromise between sensitivity and specificity making it very suitable for real monitoring systems where both missed disease and false alarm should be minimized. Impact of Adaptive Cross-Modal Attention on ROC Performance. In order to assess the impact of the proposed Adaptive Cross-Modal Attention Fusion (ACMAF) module, ROC curves were plotted for three cases: Vision Foundation Encoder only, Static Multimodal Fusion, Proposed ACMAF.

6.5 Ablation Study

Experimental setup illustrated and, quantitative results summarized in Table 13. Ablation studies quantitatively study the contribution of each major component in the proposed Trustworthy Multimodal Foundation Framework (TMFF). While performance shows that the complete framework is superior, it does not clarify how each proposed module leads to the final performance. As a result, a systematic evaluation of each module was performed by adding them one by one to the framework with identical data, training protocols, hyperparameters, and evaluation protocols. The ablation trials evaluate the effectiveness of… 1. Environmental Context Encoder (ECE). 2. Cross-modal attention fusion using adaptive methods. 3. Recognition Module for Open Worlds. 4. Reliability Evaluation Tool (5 words) The total architecture of TMFF refers to the combination of all five modules.

6.5.1 Contribution of Environmental Context Learning

Experiments were conducted through image-only learning, static multimodal fusion, and Environmental Context Encoder + ACMAF to probe the importance of multimodal information.

Table 13. Effect of environmental information.

Configuration	Accuracy (%)	F1-score (%)	AUC	MCC
Image Only	97.84	97.68	0.9918	0.975
Image + Static Environmental Fusion	98.52	98.39	0.9951	0.982
Image + ECE + ACMAF (TMFF)	99.18	99.09	0.9986	0.989

The results reveal that using environmental information greatly enhances tomato disease diagnosis. The classification accuracy was improved on 0.68% by static feature concatenation. The proposed Environmental Context Encoder and ACMAF achieved an additional improvement of 0.66%. It proves that adaptive multimodal representation learning is much more effective than the conventional method of feature fusion. The improvement is especially noted for ailments with a strong reliance on environmental conditions such as Late Blight, Leaf Mold and Target Spot.

6.5.2 Contribution of Adaptive Cross-Modal Attention Fusion (ACMAF)

Since ACMAF is the main methodological contribution of this research, we conducted further experiments to compare other multimodal fusion strategies

Table 14. Comparison of multimodal fusion strategies.

Fusion Strategy	Accuracy (%)	F1-score (%)	AUC
Feature Concatenation	98.31	98.19	0.9948
Feature Addition	98.45	98.32	0.9952
Weighted Average	98.59	98.47	0.9958
Cross-Attention	98.79	98.67	0.9967
Adaptive Cross-Modal Attention Fusion (ACMAF)	99.18	99.09	0.9986

ACMAF outperforms the traditional multimodal fusion approaches in all evaluation metrics by a large margin. In contrast to traditional fusion techniques that rely on a fixed feature importance for every modality, ACMAF dynamically adjusts the feature importance according to the disease characteristics and conditions, resulting in more discriminative multimodal representations. The suggested adaptive weighting mechanism allows the model to focus more on the visual features when disease symptoms are apparent while using more environmental context when the disease is at the early stage or visual ambiguity occurs.

6.5.3 Comparison with Existing Open-Set Recognition Methods

To further demonstrate the efficacy of OWRM, the proposed framework was compared with representative algorithms for open-set recognition for tomato disease diagnosis.

Table 15. Comparison with state-of-the-art open-set recognition methods.

Method	Known Accuracy (%)	Unknown Detection (%)	AUROC	FPR95 ↓	OSCR
SoftMax Threshold	98.64	78.35	0.892	12.43	88.72
OpenMax	98.52	88.64	0.936	7.84	93.15
Energy-Based OOD	98.61	91.82	0.957	5.61	95.18
Mahalanobis Detector	98.72	93.14	0.968	4.98	96.07
Prototype Network	98.88	95.26	0.979	3.62	97.15
TMFF (Proposed OWRM)	99.05	97.88	0.989	2.14	98.41

The findings show that the suggested OWRM outperforms existing open-set recognition techniques. Because Neural Networks make over-confident predictions for previously unseen disease samples, conventional SoftMax thresholding fails badly. Unknown disease detection was enhanced remarkably with OpenMax and Energy-Based methods but their performance is largely limited due to visual use only. In contrast, the proposed TMFF performed the best on all metrics thanks to multimodal feature learning, prototype-based representation, adaptive threshold estimation, and confidence-aware decision making. When we compared TMFF to Prototype Network which is the strongest baseline, TMFF achieves a 2.62 percentage points improvement in unknown disease detection, a 1.48 percentage points reduction in FPR95, and a 1.26 percentage points increase in OSCR.

6.5.4 Unknown Disease Detection Analysis

To conduct a better understanding of the behavior of the proposed OWRM, all unknown disease samples were classified into four groups based on their origin.

Table 16. Unknown disease detection by category.

Unknown Category	Detection Rate (%)
Nutrient Deficiency	98.24
Abiotic Stress	97.91
Mixed Disease Infection	97.36
Novel Disease Symptoms	98.02
Overall	97.88

The devised model exhibited similarly high detection performance in any unknown disease category. The best detection rate was 98.24 percent for nutrient deficiencies due to their symptom distribution being different from known disease classes. In Table 16 depicted mixed infection was the most challenging with 97.36% detection accuracy. The result is anticipated since mixed diseases show symptoms which have shared visual characteristics with multiple known diseases. However, the suggested multimodal representation successfully recognized these samples as unknown instead of assigning them to any of the diseases.

7. DISCUSSION

Experimental results show that the proposed Trustworthy Multimodal Foundation Framework (TMFF) offers a holistic framework for achieving reliable diagnosis of tomato diseases, simultaneously addressing the issues of classification accuracy, multimodal representation learning, open-world recognition, prediction reliability and computational efficiency. In contrast to traditional methods which mainly focus on optimizing classification of closed set images, TMFF integrates visual foundation representations with environmental context and reliable decision mechanisms. The framework has an high accuracy which were high values for the laboratory and real-field images for each category of tomato diseases. The high precision and recall values indicate a balance between false positive and false negative predictions, with the high AUC value showing good class separability at various decision thresholds.

The proposed framework will also have significant implications for precision agriculture. Early disease detection allows farmers and agronomists to take actions to manage disease prior to the severity of the disease. Environmental context can aid in more focused assessment of disease and can help limit the use of pesticides when they are not necessary; open-world recognition serves as an added safety factor for recognizing conditions that may not be categorized as a disease. Confidence, uncertainty and explanation information can also contribute to the transparency and facilitate informed agricultural decision making. Therefore, TMFF should be considered as an expert tool for decision-making and not as a substitute to agronomic or laboratory skills. High confidence known class predictions can be used for routine monitoring, while low confidence or unknown predictions should be referred for expert further investigation.

The results are promising, but there are some caveats. The current study does not cover streams of environmental information that are continuously synced in real-time from large-scale agricultural fields, and a more comprehensive test of the concept is thus needed, with a geographically distributed and multi-seasonal set of data. In addition, OWRM is capable of recognizing some of the unknown

samples but cannot yet automatically discover and name new disease categories. This limitation would need to be overcome by continual learning and incremental class discovery. With relatively low overheads of the proposed modules, the Vision Foundation Encoder is still computationally expensive for low-power embedded devices. Lastly, the present study is based on tomato disease diagnosis, and more study needs to be conducted in various crops, disease types, geographical regions and agricultural environments to generalize it.

Larger multi-regional and multi-season field datasets with coordinated real-time IoT measurements should be used in future studies to evaluate TMFF. The integration of imagery captured by UAVs, hyperspectral sensing, and weather forecasting may yield further information to support a full monitoring of crop health. Continual and federated learning might contribute to the adaptation to new diseases and distributed agricultural intelligence, while lightweight foundation models, agricultural vision-language models, digital twins, and autonomous agricultural robots would enhance scalability and practical deployment. In summary, the results suggest that the TMFF contributes to the tomato disease diagnosis, moving it beyond simple image classification to a multimodal open world and trustworthy AI paradigm, with its predictive success, unknown-condition detection, calibration, interpretability, and computational efficiency, to be potentially deployed within precision agriculture.

8. CONCLUSION

This article suggests a trustworthy multimodal foundation framework (TMFF) which provides smart tomato disease diagnosis and predicting early disease progression in real agro-environment. Unlike conventional methods of diagnosing plant diseases from visual symptoms, TMFF uses an unusual one. It integrates the four main modules: Vision Foundation Encoder (VFE), Environmental Context Encoder (ECE), Adaptive Cross-Modal Attention Fusion (ACMAF) and Open-World Recognition Module (OWRM) to present a comprehensive learning challenge. In this way, it offers a multimodal system perfect for open-world applications and that can provide reliable results. The integrated architecture is based on visual and contextual features, which results in an accurate diagnosis of diseases in various field conditions, along with a suitable confidence estimation and uncertainty quantification, and a suitable interpretation. Experimental outcomes on benchmark datasets and simulated real-field environments verify the superior performance of TMFF. The micro-averaged and macro-averaged area under the curve scores were 0.9918 and 0.9987, respectively, while the precision, recall, and F1 scores were 99.07%, 99.12%, and 99.09%, respectively. In addition, the proposed ACMAF module effectively improves the learning of multimodal features by dynamically modeling the interaction between visual and ambient features. The OWRM discovered previously undetected disease conditions at a rate of 97.88% and achieved Open Set Classification Rate (OSCR) of 98.41%, thus enhancing the reliability of diagnosis in real field situations. The TAM's Expected Calibration Error (ECE) was measured to be 0.012 and the Trust score was recorded at 97.42%, which shows highly calibrated, transparent, and trustworthy predictions. Our ablation studies showed that all proposed modules and particularly the ACMAF are beneficial for the improvements in accuracy classification, multimodal representation learning, and open-world recognition. In addition, at roughly 53 FPS, TMFF also allows for real-time inference which can be used in modern edge-enabled agricultural systems. In the future, we will validate the framework using extensive geographical datasets from the fields. The framework will be made more robust with IoT and UAV real-time data. Further, the framework will be further developed for different kind of crop species. The framework will also be continually updated to incorporate new data. In sum, TMFF creates a powerful new-generation framework for intelligent crop disease diagnosis, having important implications for precision agriculture and sustainable crop production.

REFERENCES.

[1] M. Brahimi, K. Boukhalfa, and A. Moussaoui, "Deep Learning for Tomato Diseases: Classification and Symptoms Visualization," Applied Artificial Intelligence, vol. 31, no. 4, pp. 299–315, 2017.

[2] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," Computational Intelligence and Neuroscience, vol. 2016, Article ID 3289801, 2016.

[3] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," Frontiers in Plant Science, vol. 7, Art. no. 1419, 2016.

[4] D. Hughes and M. Salathé, "An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics," arXiv preprint arXiv:1511.08060, 2015.
[5] A. Fuentes, S. Yoon, S. Kim, and D. S. Park, "A Robust Deep-Learning-Based Detector for Real-Time Tomato Plant Diseases and Pests Recognition," Sensors, vol. 17, no. 9, p. 2022, 2017.
[6] A. Fuentes, S. Yoon, J. Lee, and D. S. Park, "High-Performance Deep Neural Network-Based Tomato Plant Diseases and Pests Diagnosis System with Refinement Filter Bank," Frontiers in Plant Science, vol. 9, Art. no. 1162, 2018.
[7] J. Liu and X. Wang, "Tomato Diseases and Pests Detection Based on Improved YOLOv3 Convolutional Neural Network," Frontiers in Plant Science, vol. 11, Art. no. 898, 2020.
[8] M. Arsenovic, M. Karanovic, S. Sladojevic, A. Anderla, and D. Stefanovic, "Solving Current Limitations of Deep Learning Based Approaches for Plant Disease Detection," Symmetry, vol. 11, no. 7, p. 939, 2019.
[9] N. Zhang, Y. Zhao, and X. Wang, "Tomato Disease Classification Based on Multimodal Fusion Deep Learning," Agriculture, vol. 12, no. 5, 2022.
[10] Y. Majeed, M. A. Khan, et al., "Standalone Edge AI-Based Solution for Tomato Disease Detection," Smart Agricultural Technology, vol. 8, 2024.
[11] K. Joshi, A. Sharma, et al., "Precision Diagnosis of Tomato Diseases Through Deep Learning with Hybrid Data Augmentation," Current Plant Biology, vol. 41, 2025.
[12] H. Sun, X. Li, et al., "Efficient Deep Learning-Based Tomato Leaf Disease Detection Through Global and Local Feature Fusion," BMC Plant Biology, vol. 25, 2025.
[13] R. Maurya, L. Rajput, and S. Mahapatra, "RAI-Net: Tomato Plant Disease Classification Using Residual-Attention-Inception Network," IEEE Access, vol. 13, 2025.
[14] B. Sowmya and S. Guruprasad, "Deep Learning Based Plant Health Disease Detection in Tomatoes Using Inception-v4 CNN and YOLOv8," Discover Artificial Intelligence, vol. 5, 2025.
[15] H. Shehu, A. Ibrahim, et al., "Early Detection of Tomato Leaf Diseases Using Vision Transformers and Transfer Learning," Computers and Electronics in Agriculture, 2025.
[16] M. Nawaz, M. Khan, et al., "Robust Faster R-CNN with CBAM for Tomato Disease Detection," Scientific Reports, vol. 14, 2024.
[17] H. Gunasekaran, S. Rajkumar, and L. Kirubhadharsini, "Lightweight Deep Learning Models for Tomato Disease Detection: A Review," Frontiers in Plant Science, 2025.
[18] M. Jelali, "Deep Learning Networks-Based Tomato Disease and Pest Detection: A Review of Research Studies Using Real Field Datasets," Frontiers in Plant Science, vol. 15, 2024.
[19] S. Panno, A. Davino, et al., "A Review of the Most Common and Economically Important Diseases That Undermine Tomato Cultivation," Agronomy, vol. 11, 2021.
[20] J. Qiu, X. Zhang, et al., "Research on Tomato Leaf Disease Recognition Based on Improved AlexNet," Heliyon, vol. 10, 2024.
[21] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.
[22] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," ICCV, 2021.
[23] M. Caron et al., "DINOv2: Learning Robust Visual Features Without Supervision," Nature, 2024.
[24] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.
[25] X. Zhai et al., "Sigmoid Loss for Language Image Pretraining (SigLIP)," ICCV, 2023.
[26] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.
[27] G. Jocher et al., "YOLOv5," Ultralytics, GitHub Repository, 2021.
[28] Ultralytics, "YOLOv8 Documentation," 2023.
[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE TPAMI, vol. 39, no. 6, pp. 1137–1149, 2017.
[30] A. Bendale and T. Boult, "Towards Open Set Deep Networks," CVPR, 2016.
[31] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," ICLR, 2017.
[32] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult, "Toward Open Set Recognition," IEEE TPAMI, vol. 35, no. 7, pp. 1757–1772, 2013.
[33] Joy, J., Somasundaram, A., Komban, M. J., N, R., & M, K. (2026). AQ-FUSIONNET: Dual-Decomposition Attention-GRU Framework

with Adaptive Optimization for Air Quality Forecasting. Environment Conservation Journal, 27(2), 656–670.

[34] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" NeurIPS, 2017.

[35] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," ICCV, 2017.

[36] Joy, J., Devaraju, D. S., & Ramkumar, D. J. (2025). Utilizing Fuzzy-Identity-Based Encryption With Proxy-Re-Encryption For Data Sharing. Journal of Theoretical and Applied Information Technology, 103(18).

[37] J. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big Data in Smart Farming – A Review," Agricultural Systems, vol. 153, pp. 69–80, 2017.

[38] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital Twin in Industry: State-of-the-Art," IEEE Transactions on Industrial Informatics, vol. 15, no. 4, pp. 2405–2415, 2019.

[39] N. Shafiq, S. Kumar, et al., "Internet of Things for Smart Agriculture: Challenges and Future Perspectives," IEEE Access, vol. 12, 2024.

[40] FAO, The State of Food and Agriculture 2024, Food and Agriculture Organization of the United Nations, Rome, Italy, 2024.