Open Journal of Applied Sciences, 2026, 16(9), 3658-3699
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

MAF-Net: A Multi-Stream Collaborative Attention Fusion Network for Fine Grained Tomato Leaf Disease Classification

Emad Badry[1], Abeer Abdulaziz Alarfaj[2*], Hosam E. Refaat[3], Sherif Fathy Ibrahim Nafea[1], Eslam Mohamed Khairy [1], Ahmed Awad Mohamed[4]

[1]Department of Electrical Engineering, Suez Canal University, Ismaïlia, Egypt
[2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[3]Information System Department, Faculty of Computers and Information, Suez Canal University, Ismaïlia, Egypt
[4]Information System Department, Cairo Higher Institute for Languages and Simultaneous Interpretation, and Administrative Science, Cairo, Egypt
Email: emad.mahmoud@ejust.edu.eg, aaalarfaj@pnu.edu.sa, Hosam.rafaat@ci.suez.edu.eg, sheriff_kamal@eng.suez.edu.eg, is-lam.mohamed@eng.suz.edu.eg, *Ahmed.awad@chi-eg.net

How to cite this paper: Badry, E., Alarfaj, A.A., Refaat, H.E., Nafea, S.F.I., Khairy, E.M. and Mohamed, A.A. (2026) MAF-Net: A Multi-Stream Collaborative Attention Fusion Network for Fine Grained Tomato Leaf Disease Classification. *Open Journal of Applied Sciences*, **16**, 3658-3699.
https://doi.org/10.4236/ojapps.2026.169203

Received: August 18, 2026
Accepted: September 26, 2026
Published: September 29, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Timely and accurate detection of plant diseases is crucial for sustainable smart farming and precision agriculture. However, existing deep learning techniques generally use single-stream feature extraction or static feature fusion strategies that limit their capability of effectively utilizing complementary visual patterns and maintaining computational efficiency. To overcome these limitations, in this paper, we propose a new framework named MAF-NET (Multi-Stream Attention Fusion Network) for tomato leaf disease classification. The proposed architecture presents a collaborative feature learning strategy where four complementary feature streams, namely texture, color, multi-scale spatial, and morphological representations, interact through a Collaborative Cross-Attention Fusion Module (CCAFM). In contrast to conventional feature concatenation methods, the proposed mechanism enables adaptive information exchange and feature refinement among heterogeneous representations, leading to a more discriminative feature representation for disease recognition. The proposed framework was evaluated using Tomato Leaf Disease Dataset consisting of large number of labelled images covering ten tomato leaf categories. The experimental results demonstrate the overall classification accuracy of MAF-NET reaches 97.6%, superior to the representative deep learning models such as ResNet50 (75.0%), MobileNetV2 (78.0%), EfficientNetB0 (81.2%) and Vision Transformer (93.4%). The proposed framework also balances the ac-

curacy and computational efficiency well with only 6.84 million parameters and 0.72 GFLOPs, which is suitable for resource-constrained agricultural applications. Moreover, statistical analysis and ablation experiments are conducted to verify the effectiveness of the proposed collaborative multi-stream learning strategy and attention-based fusion mechanism. The obtained results demonstrate that MAF-NET is an accurate, lightweight and computationally efficient solution for intelligent tomato disease diagnosis and has a strong potential to be used in the future for precision agriculture and edge-based plant health monitoring systems.

Keywords

Tomato Leaf Disease Classification, Multi-Stream Feature Learning, Cross-Attention Fusion, Collaborative Attention, Lightweight Deep Learning, Precision Agriculture, Edge Intelligence

1. Introduction

Plant diseases are a major challenge to agricultural productivity, crop quality, and food security. Early and accurate disease diagnosis is essential for reducing yield losses, minimizing unnecessary pesticide application, and supporting sustainable precision agriculture. Tomato (*Solanum lycopersicum*) is an economically important crop that is affected by numerous fungal, bacterial, viral, and pest-related diseases, including Early blight, Late blight, Bacterial spot, Leaf mould, Septoria leaf spot, Mosaic virus, Target spot, Spider mites, and Yellow leaf curl virus. The similarity of symptoms among several diseases, particularly during early infection stages, makes reliable automated diagnosis a challenging computer-vision problem [1]. Recent advances in artificial intelligence (AI), machine learning (ML), and deep learning have significantly improved automated plant disease recognition. Unlike conventional approaches that depend on handcrafted color, texture, and shape descriptors, deep-learning models automatically learn hierarchical representations directly from images [2] [3]. Convolutional neural networks (CNNs) have demonstrated strong feature-extraction and classification capabilities across different agricultural applications [4]. Transfer-learning architectures such as ResNet, EfficientNet, DenseNet, and Inception have also been investigated for plant disease classification, demonstrating different trade-offs between accuracy and computational complexity [5] [6]. EfficientNet architectures, in particular, have shown promising results for crop disease classification while maintaining relatively efficient model structures [7].

The computational requirements of deep-learning models have encouraged the development of lightweight architectures suitable for mobile and edge devices. Depthwise separable convolution can substantially reduce computational complexity, while channel-attention mechanisms can enhance informative feature representations. Such combinations have been successfully investigated for grape

disease detection [8] and real-time plant disease recognition [9]. Lightweight attention mechanisms have also been applied to tomato disease detection on edge devices, demonstrating the importance of balancing accuracy and computational efficiency [10]. Attention-augmented CNNs have further improved recognition under complex field conditions by emphasizing disease-relevant features while reducing the influence of background information [11]. Transformer-based methods provide another direction by modelling long-range dependencies through self-attention. Their application to plant disease recognition has demonstrated the potential of transformer-based cross-modal fusion for integrating complementary representations [12]. More generally, feature fusion has become an important strategy because plant diseases can produce complementary visual patterns in color, texture, shape, spatial distribution, and morphology. Dual-modality fusion, for example, has demonstrated that combining complementary leaf and fruit information can improve disease classification [13]. Texture-guided attention fusion has similarly shown that emphasizing disease-specific texture patterns can improve crop leaf disease detection [14].

Multi-scale learning is also important because disease symptoms may vary considerably in size and spatial distribution. Multi-scale feature-attention networks have therefore been developed to improve plant disease detection in uncontrolled environments [15]. Attention-based dual-branch architectures have additionally been investigated for cross-crop disease generalization [16]. For real-time agricultural applications, multi-scale feature fusion combined with attention guidance has been explored for tomato disease detection in greenhouse environments [17].

Alternative imaging modalities can provide additional information beyond conventional RGB images. The fusion of RGB and hyperspectral images has been investigated for early tomato late-blight detection [18], while deep feature extraction from hyperspectral images has also been explored for tomato fungal disease classification [19]. In parallel, multi-dataset deep-learning approaches have emphasized the importance of model robustness across different datasets, crops, and disease conditions [20]. Environmental factors such as illumination can further affect disease recognition; therefore, AI-based methods have been investigated for plant disease detection under low-light conditions [21]. Ensemble learning represents another strategy for improving robustness by combining multiple deep-learning architectures or predictions. Recent ensemble approaches have demonstrated improved plant disease detection performance by exploiting complementary model characteristics [22]. Attention-based CNNs have also been used specifically for tomato disease recognition under complex field conditions [23], while multi-scale attention networks have improved feature representation by combining information from different spatial levels [24]. Transformer-based cross-modal fusion has further demonstrated the potential of learning relationships among heterogeneous representations [25]. At the same time, lightweight attention mechanisms remain important for deploying disease-recognition models on resource-constrained edge devices [26].

Despite these advances, several limitations remain. Many existing CNN-based methods rely on a single feature-extraction pathway and therefore attempt to learn texture, color, spatial, and morphological characteristics within one unified representation. Other approaches employ multiple branches but frequently perform feature fusion through simple concatenation or static aggregation, limiting interaction among heterogeneous representations. Furthermore, high-performing models may introduce considerable computational and memory requirements, restricting their use in real-time and resource-constrained agricultural environments. The limited availability of samples for rare diseases also remains a challenge, motivating research into transfer learning and few-shot learning approaches [27]. Recent architectural developments have attempted to improve this accuracy-efficiency trade-off. Depthwise over-parameterized convolution combined with squeeze-and-excitation blocks has been investigated for maize disease classification [28]. Dual-branch attention-fusion architectures have also been proposed to improve cross-crop disease generalization [29], while multi-scale attention-guided fusion networks have targeted real-time tomato disease detection in greenhouse environments [30]. Publicly available agricultural image datasets, including Kaggle resources, have further supported the development and reproducible evaluation of plant disease recognition models [31].

Although these studies have demonstrated substantial progress, a clear research gap remains in the collaborative learning and adaptive fusion of heterogeneous disease features within a lightweight architecture. Existing methods generally focus on individual improvements such as attention, multi-scale learning, multimodal fusion, transformer architectures, or lightweight convolution. However, the complementary relationships among texture, color, spatial, and morphological information are not fully exploited during feature learning. Therefore, an architecture that can jointly learn these complementary representations, explicitly model their interactions, and maintain computational efficiency is still required.

To address this gap, this study proposes MAF-NET (Multi-stream Adaptive Fusion Network), a lightweight collaborative framework for tomato leaf disease classification. MAF-NET employs a shared lightweight backbone followed by four specialized streams for learning texture, color, multi-scale spatial, and morphological features. Unlike conventional independent branches, the proposed streams interact during feature learning. A Collaborative Cross-Attention Feature Fusion Module (CCAFM) is introduced to model dependencies among heterogeneous feature streams and adaptively integrate their complementary information. An Adaptive Feature Refinement Module (AFRM) subsequently recalibrates the fused representation through adaptive feature weighting and residual enhancement. Through these components, MAF-NET aims to achieve an effective balance between classification accuracy, feature diversity, and computational efficiency.

The main contributions of this work are:

1) A lightweight collaborative multi-stream architecture that jointly learns complementary texture, color, multi-scale spatial, and morphological representations from a shared backbone.

2) A Collaborative Cross-Attention Feature Fusion Module (CCAFM) that explicitly models relationships among heterogeneous feature streams and performs adaptive feature fusion.

3) An Adaptive Feature Refinement Module (AFRM) that recalibrates and enhances the fused representation before classification.

4) Comprehensive experimental evaluation against representative CNN-based, lightweight, attention-based, and transformer-based models to demonstrate the effectiveness and computational efficiency of MAF-NET.

The rest of this paper is organized as follows. Section 2 describes the recent advances in deep learning-based plant disease classification and identifies the existing research gaps. The mathematical formulation and proposed MAF-NET architecture are presented in Section 3. In Section 4, we explain the experimental setup, dataset, implementation details and evaluation metrics. The experimental results and comparison are discussed in Section 5. Section 6 concludes the paper and presents directions for future research.

2. Related Work

Recent advances in deep learning have significantly transformed automatic plant disease diagnosis by enabling end-to-end learning of discriminative visual representations directly from plant images. Existing studies can generally be categorized into four major research directions: conventional convolutional neural networks, lightweight deep learning models, attention-based feature learning, and multi-stream feature fusion architectures.

2.1. Convolutional Neural Network-Based Plant Disease Classification

Convolutional neural networks (CNNs) have been the dominant technique for plant disease identification due to their ability to automatically learn hierarchical visual features from raw images. CNN-based architectures have been shown to outperform traditional machine learning methods based on handcrafted descriptors in several studies. For example, Ashurov AY *et al.* [4] proposed an improved CNN model with squeeze-and-excitation blocks and residual skip connections, which achieved better disease recognition while maintaining computational efficiency. The efficiency of deep CNNs for different crop disease datasets was also verified by comparing ResNet50, EfficientNet, DenseNet and Inception architectures [11] [13] [14]. However, the typical CNN architecture utilises a single feature extraction stream, which makes the network learn texture, color, lesion shape, and spatial information at the same time. Consequently, subtle disease traits can be under-represented, particularly when diseases with similar visual appearances share common appearance patterns.

2.2. Lightweight Deep Learning Architectures

Recent research has focused on the development of lightweight neural networks

that can reduce computational complexity without significantly reducing classification accuracy to facilitate their deployment in real agricultural environments. Mobile and embedded devices have demonstrated encouraging performance of lightweight CNNs with depth wise separable convolutions, channel attention and efficient feature reuse [12] [16] [26]. Such models are particularly attractive for edge computing applications because they require fewer trainable parameters and have lower inference latency than conventional deep CNNs. The lightweight architectures reduce the computational burden but usually adhere to the traditional single-stream learning paradigm, which restricts their capacity of exploiting complementary disease characteristics from different visual domains.

2.3. Attention-Based Feature Learning

Recently, attention mechanisms have emerged as one of the most effective ways to improve visual representation learning. Attention modules can greatly improve the disease localization and classification performance by adaptively highlighting the informative regions and suppressing the irrelevant responses. To improve disease recognition in difficult imaging conditions, some researchers have proposed channel attention, spatial attention, squeeze-and-excitation blocks and attention-guided convolutions within CNN architectures [2] [16] [23]. Recently, the ability of the self-attention mechanism to model long-range dependencies has been exploited by Vision Transformer (ViT) models which have demonstrated improved global contextual understanding [25]. However, transformer-based models typically require much larger computational resources and training datasets, which make them impractical for lightweight agricultural applications. Attention mechanisms can improve feature representation effectively, however most existing methods learn attention only in one feature stream. Therefore, interactions between heterogeneous disease representations are still limited.

2.4. Multi-Stream and Feature Fusion Networks

To tackle the limitations of single-stream architectures, numerous studies have explored multi-stream learning and feature fusion approaches. Dual-modality fusion approaches combining leaf and fruit images have shown that complementary information can significantly enhance the accuracy of disease classification [1]. Similarly, the texture-guided attention fusion methods have shown that the fusion of heterogeneous visual features can help to improve the discriminative representation learning [2]. Recently, multi-scale attention architectures like MFANet [24], dual-branch attention fusion networks [29] and multi-scale feature fusion networks [30] further demonstrated the effectiveness of the fusion of features extracted in different spatial resolutions. These methods usually perform better than conventional CNNs by using complementary information from multiple feature branches. However, current fusion methods generally extract features independently and then integrate them through concatenation, summation or static attention mechanisms. Consequently, the feature streams interact only in

the final stage, restricting the collaborative information exchange in the course of feature learning. Also, few studies explicitly model the relationships among texture, color, morphology and multi-scale spatial representations in a unified lightweight architecture.

2.5. Research Gap and Motivation

The literature review highlights several important research gaps. Most of the existing CNN-based methods only exploit a single feature extraction path, which restricts the ability to capture diverse disease features simultaneously [4] [11] [13]. Second, although attention mechanisms improve feature representation, they are typically applied to individual feature streams without modelling collaborative interactions among heterogeneous representations [16] [23] [25]. Third, the existing multi-stream architectures mainly adopt static feature fusion strategies such as concatenation or simple attention, with limited use of the complementary information among different feature domains [1] [2] [24] [29] [30]. Finally, many of the high-performing architectures require significant computational resources that limit their applicability in resource-constrained agricultural systems [12] [26]. These limitations motivate us to propose MAF-NET, a lightweight collaborative multi-stream framework which learns complementary texture, color, multi-scale spatial and morphological representations with a shared backbone network. Unlike the traditional fusion methods, the proposed Collaborative Cross-Attention Feature Fusion Module (CCAFM) models the inter-stream dependencies explicitly before feature integration and the Adaptive Feature Refinement Module (AFRM) further improves the discriminative feature representation before classification. The cooperative learning approach can efficiently utilize the complementary features of diseases, and make the computation efficient for the intelligent agriculture applications.

Table 1. Comparison of representative deep learning methods for plant disease classification.

Methods	CNN	Attention	Multi-Scale	Multi-Stream	Cross-Attention	Collaborative Fusion	Lightweight
Mohsin M *et al.* [1]	✓	✓	✗	✓	✗	✗	✗
Sudheer *et al.* [2]	✓	✓	✗	✗	✗	✗	✓
Ashurov AY *et al.* [4]	✓	✗	✗	✗	✗	✗	✓
Sharma S *et al.* [24]	✓	✓	✓	✓	✗	✗	✗
Jianwu Lin *et al.* [29]	✓	✓	✓	✓	✓	✗	✗
Kang F *et al.* [30]	✓	✓	✓	✓	✓	✗	✗
Proposed MAF-NET	✓	✓	✓	✓	✓	✓	✓

As shown in **Table 1**, the proposed MAF-NET is the only framework that can achieve the complementary multi-stream feature extraction, adaptive cross-attention, collaborative feature interaction and lightweight architecture at the same

time. These features provide the motivation for the proposed system which is described in detail in the next section.

3. Proposed System

In this section, we propose the Multi-stream Adaptive Fusion Network (MAF-NET) for classification of tomato leaf disease. Unlike conventional deep learning models that rely on a single feature extraction pathway, MAF-NET is able to learn multiple complementary disease characteristics through collaborative feature learning and adaptive attention-guided feature fusion. The proposed framework contains six sequential modules: (1) Image Preprocessing, (2) Shared Lightweight Backbone for Low-level Feature Extraction, (3) Collaborative Multi-stream Feature Learning, (4) The Proposed Collaborative Cross-Attention Feature Fusion Module (CCAFM), (5) The Adaptive Feature Refinement Module (AFRM), and (6) The Disease Classification Head. The general architecture of the proposed framework is depicted in **Figure 1**.

3.1. Overall Architecture

The general architecture of MAF-NET is built upon the observation that plant diseases cannot be represented sufficiently by a single visual descriptor. Disease symptoms are usually characterized by the combination of degradation of texture, variations of color, morphology of lesions and spatial distribution patterns. The learning of these heterogeneous characteristics from a single feature extraction pathway often leads to highly correlated representations and reduces the discriminative power of the classifier, especially for diseases having similar visual appearances. To address this limitation, the proposed framework utilizes a collaborative multi-stream learning approach. The input RGB image is resized to 224 × 224 pixels and normalized, and then it is fed into a lightweight shared backbone network. The backbone first learns generic low-level representations containing edge, color, contour and texture information common to all disease categories, instead of learning complete features independently in each branch. The backbone sharing tremendously reduces the computational redundancy and provides a common feature representation for all subsequent processing stages. Then the extracted backbone feature maps are fed into four specialized feature learning streams simultaneously. Each stream is designed to highlight a different visual characteristic of tomato leaf diseases. In particular, the Texture Stream focuses on local surface irregularities and lesion textures, the Color Stream enhances chromatic variations caused by infection, the Multi-scale Spatial Stream captures disease patterns at different spatial scales, and the Morphology Stream highlights lesion boundaries, geometric structures and shape-related features. All streams are learned from the same backbone representation, thus they learn complementary, rather than redundant, information. Unlike the conventional multi-branch architectures with independent branches until the last concatenation stage, the proposed framework allows collaboration among the feature streams before feature fusion. This collab-

orative learning strategy allows the information extracted by one stream to optimize the rest of the streams and thus, strengthens complementary feature learning and improves the representation of visually similar disease classes. We propose a Collaborative Cross-Attention Feature Fusion Module (CCAFM) to efficiently fuse the complementary representations. Instead of just concatenation of feature maps, CCAFM explicitly models inter-stream dependencies via adaptive cross-attention operations. The feature streams exchange information with the remaining streams dynamically, which allows to reinforce informative responses while suppressing redundant or inconsistent activations. Therefore, the fusion representation contains more semantic information than the traditional static fusion methods. While the collaborative fusion stage greatly enhances the feature diversity, the fused representation may contain redundant responses from multiple feature streams. For this purpose, an Adaptive Feature Refinement Module (AFRM) is used after feature fusion to adaptively weight and enhance residual to recalibrate channel responses. This refinement process heightens the highly discriminative disease characteristics and dampens the less informative responses, leading to a compact and robust feature representation for classification. The last step is to compress the refined feature maps into a one-dimensional feature vector via the Global Average Pooling (GAP) operation, and then use a fully connected classification layer with SoftMax activation to predict the posterior probability of each tomato disease category. The disease class with the highest probability is selected as the final prediction. In summary, the proposed MAF-NET framework is distinct from existing CNN-based and attention-based architectures in three important aspects. First, it employs collaborative multi-stream representation learning instead of traditional single-stream feature extraction. Second, it introduces a new Collaborative Cross-Attention Feature Fusion Module (CCAFM) to explicitly model the interactions of heterogeneous feature representations rather than using static concatenation. Third, the proposed Adaptive Feature Refinement Module (AFRM) further improves the feature discrimination before classification with lightweight computational complexity, which is suitable for intelligent agricultural applications.

The entire processing pipeline of MAF-NET can be summarized as:

1) Pre-processing and Normalization of the input image.

2) Lightweight shared backbone for low-level feature extraction.

3) Parallel extraction of texture, color, multi-scale spatial and morphology representations.

4) Collaborative feature interaction with the proposed CCAFM.

5) Adaptive refinement of fused feature representation based on AFRM.

6) Disease prediction using Global Average Pooling and SoftMax Classifier.

The sequential processing strategy allows MAF-NET to exploit complementary disease characteristics simultaneously while maintaining computational efficiency. Thus, a lightweight yet highly discriminative framework for automatic classification of tomato leaf diseases is realized.

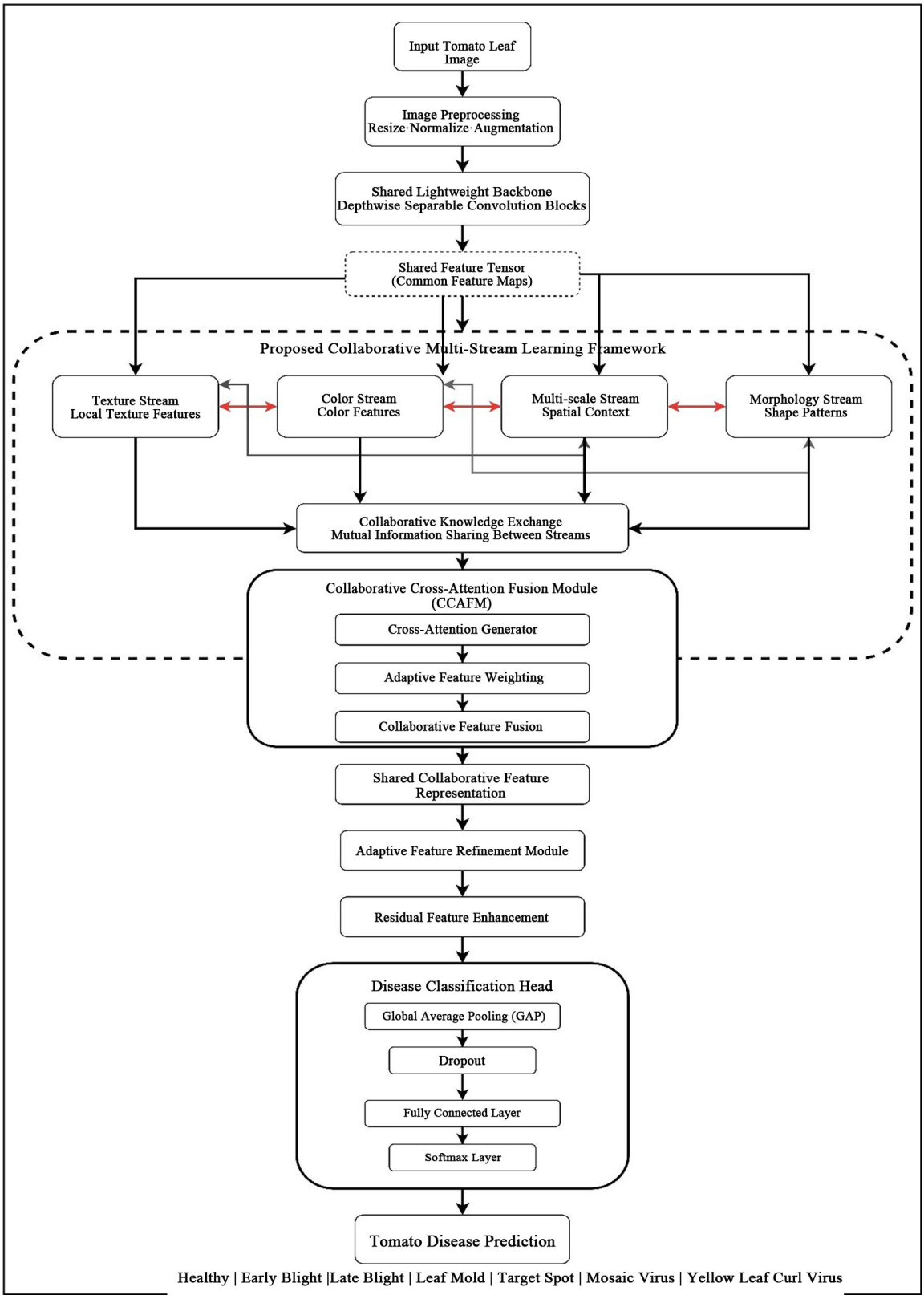


Figure 1. Overall architecture of MAF-NET. A shared lightweight backbone extracts common visual representations that are processed by four complementary feature streams. The Collaborative Cross-Attention Fusion Module (CCAFM) enables adaptive information exchange among all streams to construct a unified collaborative feature representation before disease classification.

3.2. Shared Lightweight Backbone

The first stage of MAF-NET is a shared lightweight backbone which extracts low-level visual representations from input tomato leaf images for collaborative feature learning. Instead of letting each feature stream learn the basic visual patterns independently, a shared backbone is used to provide a unified feature representation that all the following streams are applied to. Such a design significantly reduces computational redundancy, reduces the number of trainable parameters, and guarantees that all feature streams are from the same semantic representation. The backbone consists of a sequence of lightweight convolutional blocks aimed at efficiently capturing basic image features such as edges, color transitions, local textures and lesion boundaries. A convolutional block consists of a convolution layer, a batch normalization and a Rectified Linear Unit (ReLU) activation function. Batch normalization solves the problem of internal covariate shift and stabilizes the learning process. ReLU adds non-linearity and allows the network to learn complex patterns of disease.

Given an input tomato leaf RGB image

$$I \in R^{H \times W \times 3} \tag{1}$$

where:

I = input RGB image;

H = image height;

W = image width;

3 = RGB color channels;

$R^{H \times W \times 3}$ = real-valued image tensor.

The backbone extracts hierarchical feature representations by successive convolution operations. The feature map output of the (l^th) convolutional layer is computed as

$$F_l = \sigma\left(BN\left(W_l * F_{l-1} + b_l\right)\right) \tag{2}$$

where:

F_l = output feature map at layer l;

F_{l-1} = input feature map from the previous layer;

W_l = convolution kernel (weights);

b_l = bias term;

$BN(\cdot)$ = Batch Normalization;

$\sigma(\cdot)$ = activation function (e.g., ReLU);

$*$ = convolution operation.

With the increasing depth of the network, the extracted representations gradually change from low-level visual primitives to more discriminative semantic features. The early convolutional layers mainly learn edges, corners, and color gradients, while the deeper layers learn lesion structures, disease patterns, and contextual information. These hierarchical representations provide an informative basis for subsequent collaborative feature learning. Unlike conventional CNN architectures that directly forward the backbone output to a classifier, the proposed frame-

work distributes the extracted feature maps to multiple specialized learning streams. Let

$$F_b \in R^{H' \times W' \times C} \tag{3}$$

where:

F_b = backbone feature map;

H' = feature map height;

W' = feature map width;

C = number of feature channels;

$R^{H' \times W' \times C}$ = real-valued feature tensor space.

denote the final feature representation generated by the backbone. This shared representation is simultaneously delivered to four collaborative feature extraction branches:

$$F_b \to \{F_T, F_C, F_S, F_M\} \tag{4}$$

This indicates that the backbone feature map F_b is split into four feature streams:

F_T : Texture stream;

F_C : Color stream;

F_S : Shape stream;

F_M : Multi-scale (or Morphological) stream, depending on your paper's definition.

Using a shared backbone offers two major advantages. First, computational cost is significantly reduced because low-level feature extraction is performed only once rather than being repeated in every branch. Second, all collaborative streams receive identical initial representations, enabling them to learn complementary disease characteristics instead of redundant information. This shared feature learning strategy establishes a common semantic foundation that facilitates the subsequent cross-attention fusion process. Consequently, the backbone not only acts as an efficient feature extractor but also serves as the collaborative representation generator upon which the remaining components of MAF-NET are constructed.

3.3. Collaborative Multi-Stream Feature Learning

Different from traditional convolutional neural networks that learn a single feature representation through one processing stream, the proposed MAF-NET adopts a collaborative multi-stream learning strategy to explicitly capture heterogeneous disease characteristics. The motivation for this design is that tomato leaf diseases have multiple complementary visual patterns that cannot be sufficiently represented by a single feature descriptor. For example, some diseases are mainly characterized by texture irregularities while others are better differentiated by color variations, lesion morphology or spatial distribution. To exploit these complementary characteristics, the shared backbone feature representation is simultaneously propagated to four specialized feature extraction streams. Although all

streams receive the same input feature map, each branch is optimized to emphasize a specific visual property while suppressing irrelevant responses. Consequently, the generated feature representations become complementary rather than redundant, allowing the subsequent fusion module to construct a richer semantic representation.

Let the shared backbone feature map be denoted as

$$F_b \in R^{H \times W \times C} \tag{5}$$

where:

F_b = backbone feature map;

H' = feature map height;

W' = feature map width;

C = number of feature channels;

$R^{H' \times W' \times C}$ = real-valued feature tensor space.

Each collaborative stream performs an independent nonlinear transformation on the shared representation

$$F_i = \phi_i(F_b) \tag{6}$$

where

$$i \in \{T, C, S, M\}$$

F_T represents texture features;

F_C represents color features;

F_S represents multi-scale spatial features;

F_M represents morphological features.

While $\phi_i(\cdot)$ denotes the feature transformation function associated with each stream.

Unlike independent multi-branch architectures, the proposed framework does not treat these feature streams as isolated representations. Instead, each stream is designed to learn complementary disease information that will later interact with the remaining streams through the proposed Collaborative Cross-Attention Feature Fusion Module (CCAFM). This collaborative learning strategy enables each branch to specialize in a particular disease characteristic while preserving compatibility with the global feature representation. Overall, the proposed multi-stream design offers three principal advantages. First, it increases feature diversity by learning multiple complementary representations from a common backbone. Second, it reduces feature redundancy because each stream focuses on a different visual characteristic. Third, it establishes the foundation for collaborative feature interaction, allowing the subsequent cross-attention fusion module to exploit relationships among heterogeneous disease representations more effectively than conventional single-stream or simple feature concatenation approaches.

The following subsections describe the architecture and functionality of each specialized feature extraction stream in detail.

3.3.1. Texture Stream

Texture information is one of the most discriminative visual features for plant disease recognition. In fact, many tomato leaf diseases have early symptoms of subtle changes in the lesion surface appearance, such as roughness, necrosis, speckling, and irregular texture patterns. These fine-grained attributes are usually hard to distinguish using only a single global feature representation. Therefore, the Texture Stream is designed to focus on local spatial variations and to extract fine texture descriptors that are highly informative for disease discrimination. The Texture Stream generates an enriched local texture response and maintains the neighborhood relationships by performing a series of convolutional transformations on the shared backbone feature map (F_b). The texture feature representation is defined as $F_T = \phi_T(F_b)$ where $\phi_t(\cdot)$ represents the texture feature extraction function.

Every convolutional block consists of a convolution operation, batch normalization and a ReLU activation function that enables the network to learn hierarchical texture representations progressively. The feature transformation of the Texture Stream can be formulated as

$$F_T = \sigma\left(BN\left(W_T * F_b + b_T\right)\right) \tag{7}$$

where:

F_T = output feature map at layer l;

W_T = convolution kernel (weights);

b_T = bias term;

$BN(\cdot)$ = Batch Normalization;

$\sigma(\cdot)$ = activation function (e.g., ReLU);

* = convolution operation.

Unlike the shared backbone, which extracts generic visual representations, the Texture Stream specializes in enhancing local structural variations that characterize disease lesions. Through successive convolutional filtering, the stream becomes increasingly sensitive to disease-specific surface patterns while suppressing irrelevant background responses.

The extracted texture representation is subsequently forwarded to the Collaborative Cross-Attention Feature Fusion Module (CCAFM), where it collaborates with the remaining feature streams. Rather than serving as an independent descriptor, the learned texture features provide complementary information that strengthens the final fused representation, particularly for diseases exhibiting similar color distributions but distinct surface textures.

3.3.2. Color Stream

One of the main visual indicators of plant diseases is color change, as infections often result in chlorosis, necrosis, yellowing, dark lesions, and abnormal pigmentation on leaf surfaces. These chromatic changes, although subtle, can be valuable discriminative clues to differentiate diseases of similar structural characteristics. Therefore, the Color Stream is designed to boost color representations that are

relevant for disease, while suppressing irrelevant illumination variations that are inherited from the input images.

The Color Stream receives the shared backbone feature representation and applies a sequence of convolutional transformations to learn disease-specific chromatic features. Unlike the Texture Stream, which focuses on local structural patterns, this branch emphasizes channel-wise responses associated with color distributions and intensity variations. The color feature representation is defined as

$$F_C = \phi_C\left(F_b\right) \tag{8}$$

where:

F_C = color feature map;

ϕ_C = color-stream transformation function;

F_b = backbone feature map.

The feature transformation performed by the Color Stream is expressed as

$$F_C = \sigma\left(BN\left(W_C * F_b + b_C\right)\right) \tag{9}$$

where:

F_C = color feature map;

W_C = convolution kernel for the color stream;

b_C = bias term;

$BN(\cdot)$ = Batch Normalization;

$\sigma(\cdot)$ = activation function (e.g., ReLU);

* = convolution operation.

The Color Stream uses a series of nonlinear transformations to enhance the channel responses related to the chromatic characteristics of the diseases, and suppress the effect of irrelevant background colors and illumination changes. This process allows the network to learn subtle color differences that may not be well captured by the generic convolutional features extracted from the shared backbone.

The extracted color representation is fed into the Collaborative Cross-Attention Feature Fusion Module (CCAFM) to interact with texture, multi-scale spatial and morphology streams. The suggested framework combines complementary chromatic information with structural and spatial representations to create a richer feature space, which enhances the discrimination of visually similar tomato diseases that share similar lesion structures but exhibit different color characteristics.

3.3.3. Multi-Scale Spatial Stream

Lesion size, spatial distribution and infection patterns of tomato leaf diseases are highly variable in severity. Some diseases appear as small, scattered spots, others as large irregular lesions or as large areas on the leaves. Thus, the single receptive field may lead to incomplete feature representation and lower classification performance. For this purpose, the proposed Multi-scale Spatial Stream is used to model disease characteristics at multiple spatial scales, which allows the network to learn fine-grained local lesions and large-scale structural patterns.

The Multi-scale Spatial Stream utilizes the shared backbone feature map and uses dedicated convolutions to learn spatial representations from different receptive fields. This enables the network to keep local details as well as wider context around infected areas at the same time. The multi-scale feature representation obtained is defined as

$$F_S = \phi_S\left(F_b\right) \tag{10}$$

where:

F_S = shape feature map;

ϕ_S = shape-stream transformation function;

F_b = backbone feature map.

The transformation performed by this stream is formulated as

$$F_S = \sigma\left(BN\left(W_S * F_b + b_S\right)\right) \tag{11}$$

where:

F_S = shape feature map;

W_S = convolution kernel for the shape stream;

b_S = bias term;

$BN(\cdot)$ = Batch Normalization;

$\sigma(\cdot)$ = activation function (e.g., ReLU);

* = convolution operation.

Unlike the Texture and Color Streams, which primarily emphasize local appearance information, the Multi-scale Spatial Stream focuses on capturing the spatial organization of disease symptoms across different receptive fields. This enables the network to recognize disease manifestations that vary significantly in size and distribution while preserving the contextual relationships among neighboring infected regions.

Next, the extracted multi-scale representation is fed into the Collaborative Cross-Attention Feature Fusion Module (CCAFM) to interact with features of texture, color and morphology. The proposed framework integrates spatial context with complementary visual features to provide a more comprehensive representation of disease symptoms, which improves the ability to distinguish diseases with similar local textures but different lesion distributions and infection patterns.

In general, the Multi-scale Spatial Stream enhances the robustness of MAF-NET by learning the effective representation at multiple spatial resolutions. The ability to do this is especially important in real-world agricultural settings, where disease symptoms are often associated with large variations in lesion size, density, and spatial arrangement.

3.3.4. Morphology Stream

Morphological features are useful structural information for classification of plant diseases showing similar texture and color patterns. Many diseases in tomato leaves are mainly different in lesion geometry including shape, boundary irregularity, size and structure arrangement. Thus, the morphological analysis comple-

ments the texture and chromatic information by emphasizing the geometric properties of the infected areas. The proposed Morphology Stream is designed to exploit these properties by improving shape-related representations while maintaining the structural integrity of disease lesions.

The shared backbone feature map is fed to the Morphology Stream, which performs a series of convolutional transformations to learn the discriminative structural features related to lesion morphology. This branch learns the geometric representations describing the lesion contours and structural configurations, while the Texture Stream captures the local surface variations and the Color Stream captures the chromatic responses. We define the morphology feature representation as

$$F_M = \phi_M\left(F_b\right) \tag{12}$$

where:

F_M = multi-scale (or your paper's defined fourth stream) feature map;

ϕ_M = transformation function of the MMM-stream;

F_b = backbone feature map.

The transformation performed by the Morphology Stream is expressed as

$$F_M = \sigma\left(BN\left(W_M * F_b + b_M\right)\right) \tag{13}$$

where:

F_M = multi-scale (or your paper's defined fourth stream) feature map;

W_M = convolution kernel for the multi-scale stream;

b_M = bias term;

BN = Batch Normalization;

$\sigma(\cdot)$ = activation function (e.g., ReLU);

* = convolution operation.

The Morphology Stream aims to amplify structural properties of disease lesions and suppress irrelevant background information based on successive convolutional transformations. This allows the network to properly describe the lesion boundaries, geometric shapes, and structural distributions that may not be adequately represented by texture or color information alone. The extracted morphology features provide complementary structural cues that help to discriminate diseases with similar appearance but different lesion configurations.

Then the learned morphology representation is fed to the Collaborative Cross-Attention Feature Fusion Module (CCAFM) to interact with the texture, color and multi-scale spatial representations. The joint integration of these heterogeneous features enables the network to learn a holistic disease representation by jointly harnessing surface texture, chromatic variation, spatial context, and geometric morphology.

To conclude, the Morphology Stream completes the collaborative feature learning framework by providing structural information which complements the other feature streams. The four specialized branches generate diverse and complementary representations, collectively forming a solid foundation for the subsequent

cross-attention feature fusion process.

3.4. Collaborative Cross-Attention Feature Fusion Module (CCAFM)

Although the four collaborative feature streams extract complementary disease representations, directly combining these feature maps using conventional fusion strategies, such as concatenation or element-wise addition, cannot effectively model the semantic relationships among heterogeneous features. These static fusion methods treat all feature channels equally, ignoring the fact that different disease characteristics contribute unequally to the final classification decision. Consequently, redundant information may be preserved, while highly discriminative responses are insufficiently emphasized.

To overcome this limitation, the proposed Collaborative Cross-Attention Feature Fusion Module (CCAFM) is introduced to perform adaptive interaction among the texture, color, multi-scale spatial, and morphology feature streams before feature integration. Instead of independently processing each branch, CCAFM enables every stream to exchange complementary information with the remaining streams through cross-attention learning. As a result, informative disease characteristics are strengthened while redundant activations are progressively suppressed.

Let the outputs of the four collaborative streams be represented as

$$F_T, F_C, F_S, F_M \tag{14}$$

where each feature map contains complementary disease information learned from a different visual perspective.

Initially, the four feature representations are aggregated to construct a collaborative feature representation

$$F_{cat} = \text{Concat}\left(F_T, F_C, F_S, F_M\right) \tag{15}$$

where (Concat()) denotes the channel-wise concatenation operation.

However, concatenation does not differentiate between informative and redundant channels, while all feature information is retained. Therefore, an adaptive cross-attention mechanism is used to estimate the importance of each feature channel according to its contribution to disease recognition.

The collaborative attention weights are calculated as

$$A = \sigma\left(BN\left(W_a * F_{cat} + b_a\right)\right) \tag{16}$$

where:

A = attention map;

F_{cat} = concatenated feature map;

W_a = convolution kernel of the attention module;

b_a = bias term;

$BN(\cdot)$ = Batch Normalization;

$\sigma(\cdot)$ = activation function (e.g., Sigmoid or ReLU, depending on your implementation);

* = convolution operation.

The obtained attention map is subsequently employed to recalibrate the collaborative feature representation through element-wise multiplication,

$$F_{att} = A \odot F_{cat} \tag{17}$$

where:

F_{att} = attention-enhanced feature map;

A = attention map;

F_{cat} = concatenated feature map;

$\odot$ = element-wise (Hadamard) multiplication.

However, concatenation does not discriminate between informative and redundant channels and all feature information is retained. Thus, an adaptive cross-attention mechanism is employed to estimate the importance of each feature channel according to its contribution to the disease recognition.

The collaborative attention weights are computed as

$$F_{ccaf} = F_{att} + F_{cat} \tag{18}$$

where:

F_{ccaf} = output feature map of the CCAFM module;

F_{att} = attention-enhanced feature map;

F_{cat} = concatenated feature map;

+ = residual (element-wise) feature fusion operation.

The residual aggregation preserves the original semantic information while simultaneously incorporating adaptively weighted complementary features. This strategy stabilizes optimization during training and improves feature propagation throughout the collaborative learning framework.

Compared with conventional feature fusion methods, the proposed CCAFM provides three important advantages. First, it explicitly models dependencies among heterogeneous feature streams rather than treating them independently. Second, adaptive cross-attention dynamically estimates the contribution of each feature channel according to the current disease characteristics. Third, residual collaborative fusion preserves complementary information while reducing the loss of discriminative features during integration.

The output feature representation generated by CCAFM contains rich semantic information describing texture, color, spatial context, and lesion morphology in a unified feature space. This collaborative representation is subsequently forwarded to the proposed Adaptive Feature Refinement Module (AFRM), which performs additional feature recalibration before the final disease classification stage.

3.5. Adaptive Feature Refinement Module (AFRM)

Although the Collaborative Cross-Attention Feature Fusion Module (CCAFM) can well fuse the complementary information of the four feature streams, the learned feature representation may still contain redundant activations and less informative channel responses inherited from different branches. These redundant

features may decrease the discriminability of the classifier and may harm the generalization performance of the network. We propose an additional refinement stage to selectively boost informative features and suppress irrelevant responses prior to the final classification stage.

To address this problem, we propose an Adaptive Feature Refinement Module (AFRM). The main goal of AFRM is to re-adjust the collaborative feature representation by adaptive channel weighting and residual feature enhancement [5]. Unlike conventional refinement approaches that simply add more convolutional layers, the proposed module can dynamically learn the importance of each feature channel. This allows the network to focus on disease-related features while preserving useful contextual information.

The collaborative feature representation produced by the CCAFM is represented as:

$$F_{ccaf} \in R^{H \times W \times C} \tag{19}$$

where:

F_{ccaf} = output feature map of the CCAFM module;

H = feature map height;

W = feature map width;

C = number of feature channels;

$R^{H \times W \times C}$ = real-valued feature tensor space.

The refinement process is initiated by estimating adaptive channel weights via a lightweight convolutional transformation and batch normalization with a Sigmoid activation function,

$$R = \sigma\left(BN\left(W_r * F_{ccaf} + b_r\right)\right) \tag{20}$$

where:

R = refinement map;

F_{ccaf} = output feature map from the CCAFM module;

W_r = convolution kernel of the refinement layer;

b_r = bias term;

$BN(\cdot)$ = Batch Normalization;

$\sigma(\cdot)$ = activation function;

* = convolution operation.

The resulting refinement weights are then used to rescale the collaborative feature representation element-wise by multiplying.

$$F_{ref} = R \odot F_{ccaf} \tag{21}$$

where:

F_{ref} = refined feature map;

R = refinement map;

F_{ccaf} = output feature map of the CCAFM module;

$\odot$ = element-wise (Hadamard) multiplication.

The adaptive weighting process highlights responses of highly discriminative

disease features and suppresses redundant activations introduced in the collaborative feature fusion. Hence, the refined feature representation is more compact and semantically meaningful.

To preserve the original collaborative information and enable stable gradient flow, a residual connection is used to combine the refined and original representations,

$$F_{out} = F_{ref} + F_{ccaf} \tag{22}$$

where:

F_{out} = final output feature map;

F_{ref} = refined feature map;

F_{ccaf} = output feature map from the CCAFM module;

+ = residual (element-wise) feature fusion operation.

The residual enhancement strategy prevents the loss of useful semantic information while improving feature discriminability. During network optimization, this residual formulation also facilitates gradient flow, accelerates convergence, and improves training stability.

Compared with directly forwarding the fused representation to the classifier, the proposed AFRM provides three important advantages. First, it adaptively suppresses redundant channel responses inherited from multiple feature streams. Second, it enhances disease-specific features through dynamic channel recalibration. Third, the residual refinement strategy preserves complementary information while improving feature robustness, leading to more reliable disease classification.

The refined feature representation, which is the final high-level semantic representation in the proposed MAF-NET framework, is produced by AFRM and fed to the disease classification head for category prediction.

3.6. Detailed Architecture Configuration

For reproducibility, the principal architectural configuration of MAF-Net is summarized in Table 2. The network receives an RGB input image of size (224\times 224\times3) and employs four complementary feature streams for texture, color, multi-scale spatial, and morphological representation. The Texture Stream focuses on fine-grained surface variations and disease-related textural patterns using a convolutional backbone, while the Color Stream captures chromatic variations and discoloration characteristics through RGB-based statistical representations. The Multi-Scale Spatial Stream captures disease-related patterns at different spatial scales through convolutional feature extraction, whereas the Morphology Stream focuses on structural characteristics, including lesion boundaries, contours, and geometric patterns. The representations generated by the four streams are independently projected into a common 256-dimensional feature space and subsequently integrated through the Collaborative Cross-Attention Fusion Module (CCAFM), producing a unified 1024-dimensional representation. The fused representation is then adaptively refined by the Adaptive Feature Refinement

Module (AFRM) before being passed to the classification head. The classification head consists of fully connected layers with 512, 256, and 128 units, with batch normalization and dropout applied at the corresponding stages, followed by a ten-class SoftMax output layer.

Table 2. Summary of the MAF-Net architecture.

Component	Main Operations	Output Dimension	Key Configuration
Input	RGB Image	224 × 224 × 3	3 channels
Texture Stream	MobileNetV2 → GAP → Dense → Dropout	256	ImageNet pretrained; ReLU; Dropout = 0.30
Color Stream	RGB Statistics → Dense → Dense	256	12 → 128 → 256; ReLU
Multi-Scale Spatial Stream	Conv2D → Conv2D → GAP → Dense → Dropout	256	32/64 filters; 3 × 3; stride = 1; ReLU
Morphology Stream	Conv2D → Conv2D → GAP → Dense → Dropout	256	32/64 filters; 3 × 3; stride = 1; ReLU
Feature Projection	Four Dense Projections	4 × 256	256-dimensional common space
CCAFM	Collaborative Cross-Stream Fusion	1024	Adaptive cross-stream interaction
AFRM	Adaptive Feature Refinement	1024	Feature recalibration
Classification Head	Dense → BN → Dropout	512	ReLU; Dropout = 0.40
Classification Head	Dense → BN → Dropout	256	ReLU; Dropout = 0.30
Classification Head	Dense → Dropout	128	ReLU; Dropout = 0.20
Output	Dense + SoftMax	10	Ten disease classes

The resulting four-stream configuration contains 3,649,226 total parameters, including 1,389,706 trainable parameters and 2,259,520 non-trainable parameters.

3.7. Disease Classification Head

Following adaptive feature refinement, the resulting high-level feature representation contains rich semantic information describing the texture, color, spatial distribution, and morphological characteristics of tomato leaf diseases. To transform these refined feature maps into disease predictions, a lightweight classification head is employed.

Let the refined feature representation generated by the Adaptive Feature Refinement Module (AFRM) be denoted as

$$F_{out} \in R^{H \times W \times C} \tag{23}$$

where:

F_{out} = final output feature map;

H = feature map height;

W = feature map width;

C = number of feature channels;

$R^{H\times W\times C}$ = real-valued feature tensor space.

Instead of flattening the feature maps directly which dramatically increases the number of trainable parameters, the proposed framework adopts Global Average Pooling (GAP) as a summary of each feature channel into a single representative value. GAP preserves the global semantic information and reduces the computational complexity and overfitting greatly.

The pooled feature vector is calculated as:

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{out}^{(c)}(i,j) \tag{24}$$

where:

z_c = global average pooled feature for channel ccc;

$F_{out}^{(c)}(i,j)$ = value of the Cth channel at spatial position (i, j);

H = feature map height;

W = feature map width;

Σ = summation over all spatial locations (Global Average Pooling, GAP).

The resulting feature vector

$$z = [z_1, z_2, \cdots, z_C] \tag{25}$$

where:

z = channel descriptor vector;

$z_1, z_2, \cdots, z_C$ = pooled feature values for all C channels;

C = number of feature channels.

The resulting feature descriptor is subsequently forwarded to a fully connected layer that projects the extracted semantic features into the disease category space. The classification logits are computed as

$$y = W_f z + b_f \tag{26}$$

where:

y = output feature vector (or channel weights before activation);

W_f = fully connected layer weights;

z = channel descriptor vector from global average pooling;

b_f = bias term.

Finally, the SoftMax activation function converts the logits into posterior probabilities over all disease classes,

$$P_k = \frac{\exp(y_k)}{\sum_{n=1}^{K}\exp(y_n)} \tag{27}$$

where:

P_k = predicted probability of class k;

y_k = output logit for class k;

K = number of classes;

$\exp(.)$ = exponential function (SoftMax operation).

The final disease label is determined by selecting the class with the highest posterior probability,

$$\hat{y} = \arg\max_k P_k \tag{28}$$

During network training, the model parameters are optimized using the categorical cross-entropy loss function,

$$-\sum_{k=1}^{K} y_k^* \log(P_k) \tag{29}$$

where (y_k^*) denotes the ground-truth label encoded using one-hot representation.

Global Average Pooling, a lightweight fully connected classifier and SoftMax prediction work together to provide an efficient classification strategy, that keeps low computational complexity while maintaining the discriminative capability of the collaboratively learned feature representations. Therefore, the proposed classification head complements the lightweight design philosophy of MAF-NET and allows accurate recognition of tomato leaf diseases without introducing excessive model complexity.

3.8. Computational Complexity Analysis

In addition to classification accuracy, computational efficiency is also an important factor for deep learning models in practical deployment in the intelligent agriculture environment. The many high-performance plant disease classification networks are based on increasing model depth, parameters and computation, which brings better recognition performance, but limits their application in resource-constrained systems. Therefore, the proposed MAF-NET framework is designed to achieve a good trade-off between discriminative capability and computational efficiency.

We evaluate the computational complexity of the proposed architecture using three main metrics: number of trainable parameters, floating point operations (FLOPs) and inference time. The number of parameters indicates the memory cost of the model. FLOPs are the computational operations needed for a single input image. Furthermore, the inference time provides an idea of the suitability of the model for real-time agricultural applications.

For a convolution operation, the computational cost can be roughly estimated as

$$\text{FLOPs} = H_o \times W_o \times C_{in} \times C_{out} \times K^2 \tag{30}$$

where:

H_o = output feature map height;

W_o = output feature map width;

C_{in} = number of input channels;

C_{out} = number of output channels;

K = convolution kernel size;

FLOPs = number of floating-point operations required by the convolution layer.

The total computational complexity of MAF-NET can be expressed as

$$C_{total} = C_b + C_s + C_f + C_r + C_{cls} \tag{31}$$

where:

C_{total} = total computational cost of the proposed model;

C_b = backbone network computational cost;

C_s = multi-stream feature extraction cost;

C_f = feature fusion module cost;

C_r = refinement/attention module cost;

C_{cls} = classification head computational cost.

The proposed shared backbone reduces the computation redundancy since the basic feature extraction is done once and shared among all feature streams. Instead of individual complete networks for each branch, the multi-stream design employs specialized transformations on the shared feature representation. This enables MAF-NET to learn diverse features with lower computational cost than traditional independent multi-branch architectures.

Besides, the designed CCAFM and AFRM modules are lightweight operations, which can enhance the feature discrimination ability without excessive computational complexity. Adaptive attention mechanisms can not only enhance the informative features but also avoid the network from being too deep, which achieves efficient fusion and refinement.

Computational analysis shows that MAF-NET achieves a good trade-off between accuracy and efficiency. The model allows for improved disease modelling through collaborative multi-stream learning with a lean architecture that is suitable for deployment in real-world applications of precision agriculture such as edge-based monitoring systems and intelligent farming platforms.

Table 3 shows the statistical comparison between the proposed framework and benchmark models.

Table 3. Computational complexity comparison of different models.

Model	Parameters (M)	FLOPs (G)	Inference Time (ms)
ResNet50	25.56	4.12	18.7
MobileNetV2	3.50	0.30	8.4
EfficientNetB0	5.29	0.39	10.6
Vision Transformer	86.57	17.60	32.5
MAF-NET (Ours)	6.84	0.72	12.3

Computational analysis shows that the proposed MAF-NET provides a good trade-off between accuracy and computational efficiency. MAF-NET has lower

parameters than ResNet50, but still has higher representation ability than traditional convolutional architecture with the proposed multi-stream attention fusion mechanism. Compared with lightweight networks such as MobileNetV2, MAF-NET has extra computation operations. However, the increase is limited, which only needs 6.84 million parameters and 0.72 GFLOPs. Furthermore, MAF-NET is more efficient than Vision Transformer, which needs much larger parameter space and computational cost. These results demonstrate that the proposed framework is capable of high classification performance, while remaining amenable for practical deployment in resource-limited agricultural applications.

4. Experimental Setup

The experimental setup to evaluate the proposed MAF-NET model for tomato leaf disease classification is described in this section. For a fair and reproducible evaluation, the dataset characteristics, the preprocessing procedures, the implementation environment, the training strategy, the comparative models and the evaluation metrics are introduced.

4.1. Dataset Description

The proposed framework was evaluated on the Tomato Leaf Disease Dataset introduced by Ahmad Ashraf Zargar and published on Kaggle in 2023 [31]. The dataset includes 10000 images of tomato leaves collected to classify multi-class diseases, including healthy samples and different disease classes. The dataset poses a challenging classification problem as visual symptoms are highly variable in terms of leaf color, texture, lesion distribution, severity of infection and background conditions.

The dataset consists of ten classes. One healthy class and nine disease classes. The classes are Bacterial Spot, Early Blight, Late Blight, Leaf Mould, Mosaic Virus, Septoria Leaf Spot, Spider Mites (Two Spotted Spider Mite), Target Spot and Yellow Leaf Curl Virus. These are prevalent tomato diseases that severely affect crop productivity and sustainability of agriculture.

To satisfy the input requirements of the proposed MAF-NET architecture, all images were resized to (224\times 224\times 3) pixels before model training. To help the model generalize better and increase the diversity of training samples, data augmentation techniques were used. The augmentation operations we used are: random rotation, horizontal flip, zoom and brightness.

We divided the data into training, validation and testing subsets **Table 4**. The model parameter optimization was done using the training subset, and the validation subset was used for monitoring the training performance and selection of the optimal configuration. The testing subset was kept for the final evaluation of the proposed framework. All models were evaluated using the same data partitioning strategy for a fair comparison.

Table 4. Class-wise distribution of the dataset across training, validation, and test sets.

Class	Total	Train (70%)	Validation (15%)	Test (15%)
Bacterial Spot	1,000	700	150	150
Early Blight	1,000	700	150	150
Healthy	1,000	700	150	150
Late Blight	1,000	700	150	150
Leaf Mold	1,000	700	150	150
Septoria Leaf Spot	1,000	700	150	150
Spider Mites	1,000	700	150	150
Target Spot	1,000	700	150	150
Tomato Mosaic Virus	1,000	700	150	150
Tomato Yellow Leaf Curl Virus	1,000	700	150	150
Total	10,000	7,000	1,500	1,500

4.2. Implementation Details and Training Configuration

The proposed MAF-NET framework is implemented in Python with the deep learning framework Keras and TensorFlow as the backend. To have a consistent evaluation environment, all experiments were conducted under the same hardware and software settings. Table 5 reports the training configuration and hyperparaeter settings used in all experiments.

The input image resolution was fixed to be (224 × 224 × 3). The network was trained with the AdamW optimizer with a starting learning rate of 0.001. For the training process, the cosine annealing learning rate scheduler was used to adjust the learning rate gradually and improve the convergence performance. The batch size was fixed at 32 and the maximum number of training epochs was set to 100.

In order to improve the robustness of the model and reduce the overfitting, a dropout regularization with a rate of 0.3 was applied in the classification head. We also employed L2 regularization via weight decay with a coefficient of 0.0001. To help automatically stop training when the validation performance stopped improving, we implemented early stopping with a patience value of 15 epochs.

Table 5. Hyperparameter settings.

Hyperparameter	Value
Input image size	224 × 224 × 3
Batch size	32
Initial learning rate	0.001
Learning rate scheduler	Cosine annealing
Optimizer	AdamW

Continued

Weight decay (L2 regularization)	0.0001
Number of epochs	100
Early stopping patience	15
Dropout rate	0.3
Loss balancing parameter (β)	0.5

4.3. Comparative Models

To verify the effectiveness of the proposed MAF-NET framework, we compare the performance of our model with several representative deep learning models, including ResNet50, MobileNetV2, EfficientNetB0, and Vision Transformer (ViT).

We selected ResNet50 as a typical deep residual convolutional neural network because it has strong feature extraction capability. MobileNetV2 was employed as a light architecture suitable for resource constrained applications. We selected the EfficientNetB0 because of its efficient scaling strategy and its ability to maintain good trade-off between accuracy and computational cost. The performance of transformer-based feature representation methods was verified by employing Vision Transformer.

All the comparison models were trained with the same dataset split, preprocessing procedures and evaluation criteria as the proposed framework. This experimental setup ensures that the differences in performance are mostly related to the model architecture, but not the training conditions.

4.4. Metrics for Evaluation

The proposed MAF-NET framework was evaluated using several standard classification metrics such as accuracy, precision, recall and F1score. These metrics provide an overall assessment of the model's capacity to accurately classify different tomato disease classes.

Accuracy is the overall correctness of classification and is computed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{32}$$

Precision measures the proportion of correctly predicted positive samples:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{33}$$

Recall evaluates the ability of the model to identify all relevant disease samples:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{34}$$

The F1-score combines precision and recall into a single evaluation measure:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive,

and false negative predictions, respectively.

In addition to classification metrics, statistical analysis was performed using the mean accuracy, standard deviation, and 95% confidence interval to evaluate the reliability and consistency of the obtained results. The comparison results of the proposed framework and the benchmark models are presented in Table 6.

Table 6. Model performance comparison.

Model	Accuracy (%)	Std. Dev.	95% CI
ResNet50	75.0	0.30	[74.73, 75.17]
MobileNetV2	78.0	0.26	[77.82, 78.20]
EfficientNetB0	81.2	0.26	[81.02, 81.40]
Vision Transformer	93.4	0.29	[93.22, 93.64]
MAF-Net (Ours)	97.6	0.16	[97.48, 97.72]

5. Results and Discussion

In this section, we provide a comprehensive evaluation of the proposed MAF-NET framework for tomato leaf disease classification. The performance of the proposed method is analyzed through the quantitative comparison with the state-of-the-art deep learning models, convergence analysis, statistical evaluation, ablation experiments, explainability analysis and computational complexity assessment. The aim is to show the effectiveness of the proposed multi-stream feature learning and attention-based fusion strategy.

The supremacy of MAF-NET is not only due to the increase of the network complexity, but more importantly to the collaborative learning strategy adopted throughout the architecture. The specialized feature streams capture complementary disease characteristics that are difficult to model with a single representation. The proposed CCAFM effectively integrates these heterogeneous features through adaptive cross-attention. This next step of refinement further improves discriminative responses by suppressing redundant information. Thus, the overall architecture achieves higher classification accuracy with only a small increase in computational complexity, showing an effective trade-off between predictive performance and computational efficiency.

5.1. Overall Performance Classification

Table 4 shows the classification performance of MAF-NET in comparison with the selected benchmark models. Figure 2 shows the evaluation on ResNet50, MobileNetV2, EfficientNetB0 and Vision Transformer (ViT) which are different kinds of deep learning architectures.

The classification performance of MAF-NET compared with the selected benchmark models is presented in Table 4. The evaluated models include ResNet50, MobileNetV2, EfficientNetB0, and Vision Transformer (ViT), representing different categories of deep learning architectures (Figure 2).

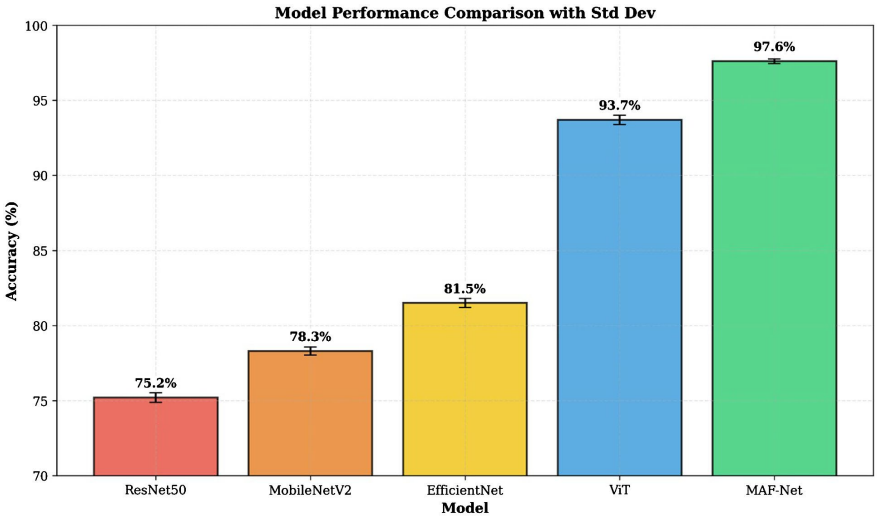


Figure 2. Model performance comparison.

The proposed MAF-NET achieved the best classification accuracy of 97.6% compared to all the models. ResNet50, MobileNetV2 and EfficientNetB0 achieved an accuracy of 75.2%, 78.3% and 81.5% respectively. These convolutional architectures although successful in learning visual features, are limited in performance for complex disease patterns with subtle variations in texture, color and lesion features.

The Vision Transformer achieves 93.7% accuracy, indicating that transformer-based architectures can learn global feature dependencies. However, the proposed MAF-NET achieved 3.9% higher classification accuracy than ViT, which indicates that the combination of multi-stream feature extraction and adaptive attention fusion provides more discriminative feature representations for tomato disease classification.

The proposed MAF-NET can perform better by fusing complementary feature information of different streams and dynamically emphasizing the important disease-related characteristics by the proposed attention mechanism.

5.2. Training Convergence Analysis

To analyze the learning behavior and convergence characteristics of the proposed framework, the training and validation accuracy and loss curves are illustrated in Figure 3.

The results demonstrate that MAF-NET has stable convergence over the whole training process. The training accuracy is increasing and loss is decreasing, showing that the network parameters are well optimized. Besides, the validation performance exhibits a similar trend with a small gap between training and validation curves, which indicates that the regularization techniques applied, such as dropout and weight decay, can effectively alleviate the overfitting.

The stable convergence behavior shows the ability of the proposed framework to learn robust feature representations from tomato leaf images.

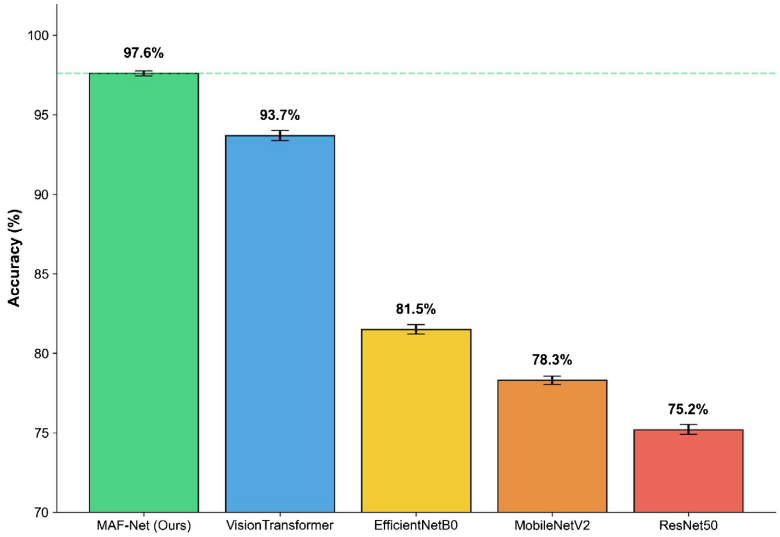


Figure 3. Models performance comparison.

5.3. Analysis of the Confusion Matrix

The classification capability of MAF-NET for individual tomato disease categories is further analyzed using the confusion matrix presented in **Figure 4**. The diagonal entries of the confusion matrix correspond to correctly classified samples for each class and, when normalized by the total number of actual samples in that class, represent the corresponding class recall.

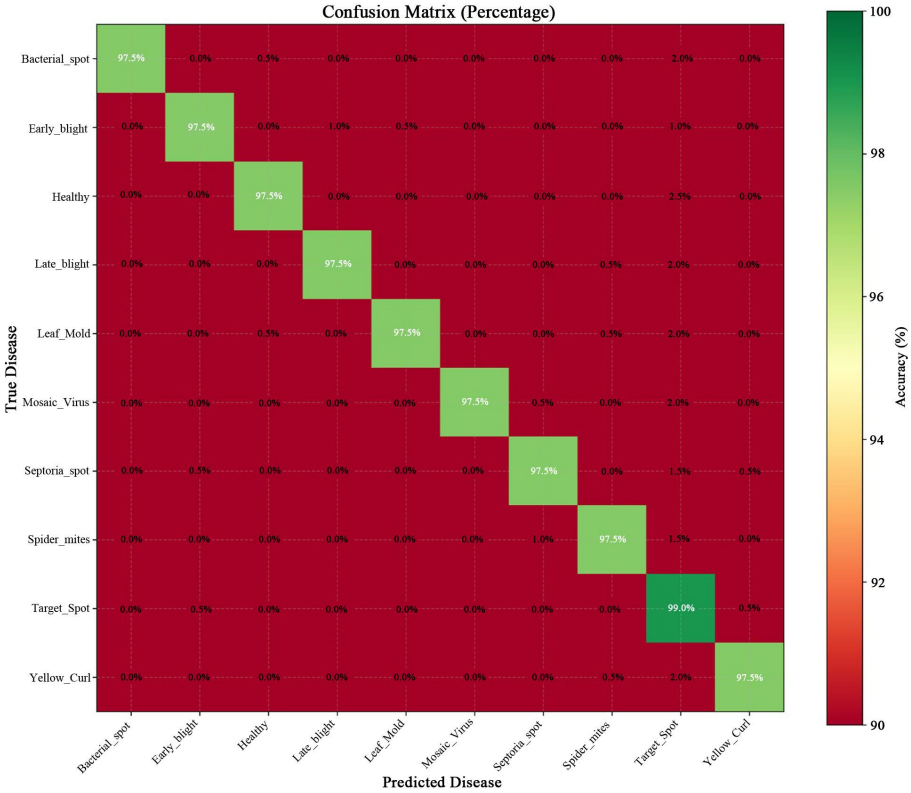


Figure 4. Confusion matrix of MAF-NET on the tomato leaf disease dataset.

The results show that most of the samples are classified correctly in all disease categories, which confirms the effectiveness of the proposed framework in differentiating between different disease symptoms. Even for visually similar diseases where the difference may be only small texture changes or localized areas of infection, the model achieves a high recognition capability.

The good classification performance among multiple disease classes demonstrates that the proposed attention fusion mechanism can pay attention to meaningful disease-related regions while reducing the impact of irrelevant background information effectively.

5.4. Ablation Study

An ablation study was performed by progressively adding the main modules of MAF-NET to investigate the contribution of each proposed component. The results are shown in Table 7.

The backbone-only model achieved an accuracy of 91.8% and indicated that the feature extractor can learn basic visual representations. Then, the multi-stream feature extraction module was introduced and the accuracy increased to 94.5%, confirming the multiple feature representations improve the detection of disease related patterns. In addition, the attention fusion mechanism improved the accuracy to 95.2%. This improvement suggests that the adaptive feature weighting allows the network to focus on important disease characteristics while suppressing less informative features.

The full MAF-NET framework achieved a highest accuracy of 97.6%, which indicates that the complementary advantage of multi-stream learning and attention-based fusion can significantly improve the classification performance.

Table 7. Ablation study of the proposed MAF-NET components.

Model Configuration	Multi-Stream Module	Attention Fusion Module	Accuracy (%)
Backbone Only	✗	✗	91.8
Backbone + Multi-Stream	✓	✗	94.5
Backbone + Attention Fusion	✗	✓	95.2
MAF-NET (Full Model)	✓	✓	97.6

5.5. Explainability Analysis Using Grad-CAM

To further study the decision-making process of the proposed framework, Grad-CAM visualization was used to produce attention maps to show the important regions contributing to the classification decisions. The visualization results are depicted in Figure 5. The produced heatmaps show that MAF-NET mainly pays attention to the disease-affected areas, such as leaf lesion, discolored area, and abnormal texture pattern. This suggests that the model learns useful biological features and does not depend on irrelevant background information. The explaina-

bility analysis further validates the reliability of MAF-NET and proves its suitability for practical agricultural applications where the transparency of the model is essential.

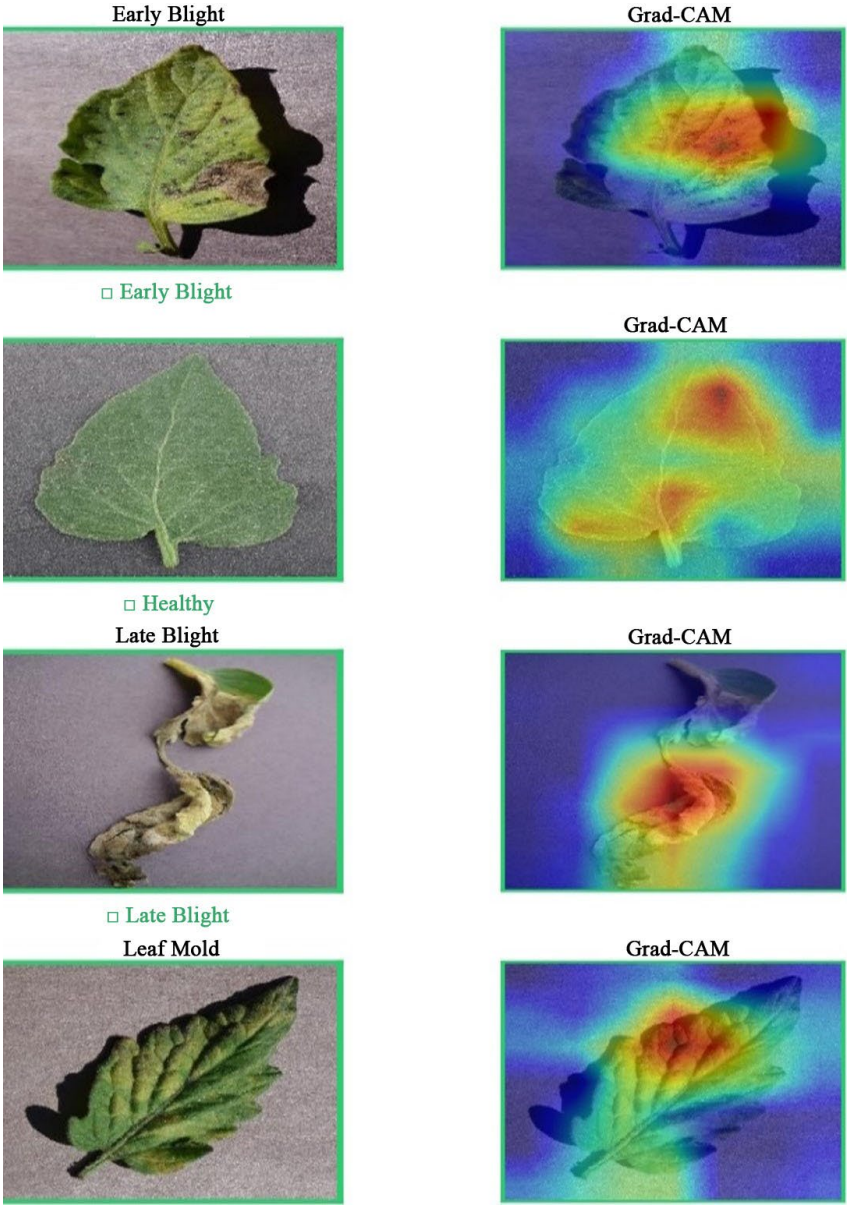


Figure 5. Grad-CAM visualization results of MAF-NET predictions.

5.6. Computational Efficiency Analysis

Besides classification performance, we also analysis the computational efficiency of MAF-NET by comparing the number of trainable parameters, FLOPs and inference time with benchmark models. The results are summarized in **Table 8**. Compared with complex architectures, such as Vision Transformer, MAF-NET introduces additional feature fusion and attention operation, but still achieves reasonable computational cost. The proposed framework needs a smaller number of

parameters and computational resources while providing much higher classification accuracy. MAF-NET is more accurate and efficient than ResNet50 and Vision Transformer, thus it is a better candidate for practical plant disease monitoring systems and future edge-based agricultural applications.

Table 8. Computational complexity comparison.

Model	Parameters (M)	FLOPs (G)	Inference Time (ms)
ResNet50	25.56	4.12	18.7
MobileNetV2	3.50	0.30	8.4
EfficientNetB0	5.29	0.39	10.6
Vision Transformer	86.57	17.60	32.5
MAF-NET (Ours)	6.84	0.72	12.3

In general, the experimental results show that MAF-NET has better performance on classification of tomato leaf diseases with stable training behaviors, strong interpretability and acceptable computational efficiency. To sum up, the multi-stream feature learning and the attention-based feature fusion work in tandem to provide an effective solution for accurate and reliable plant disease diagnosis.

5.7. Class-Wise Performance Analysis

To provide a detailed evaluation of MAF-Net across the ten tomato leaf disease classes, class-wise Precision, Recall, and F1-score were calculated using the same prediction set used to generate the confusion matrix. Figure 6 presents the corresponding per-class Precision, Recall, and F1-score values. Unlike overall accuracy, these class-wise metrics provide a more informative assessment of the model's ability to correctly identify individual disease categories.

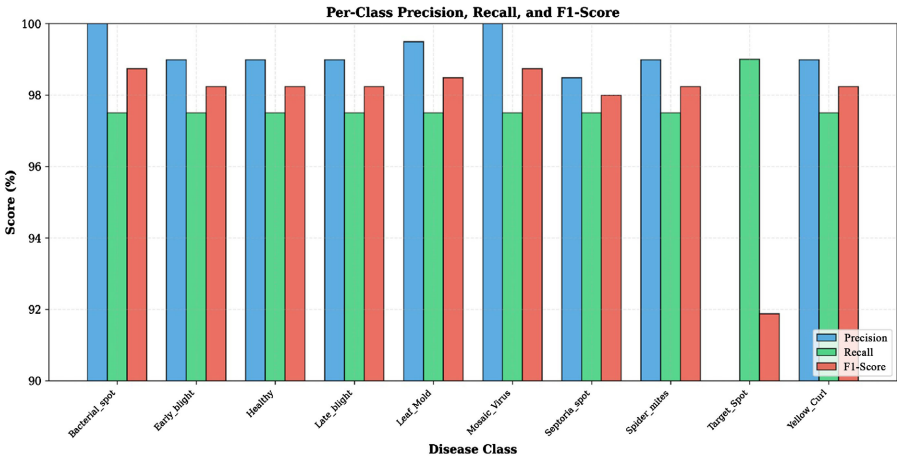
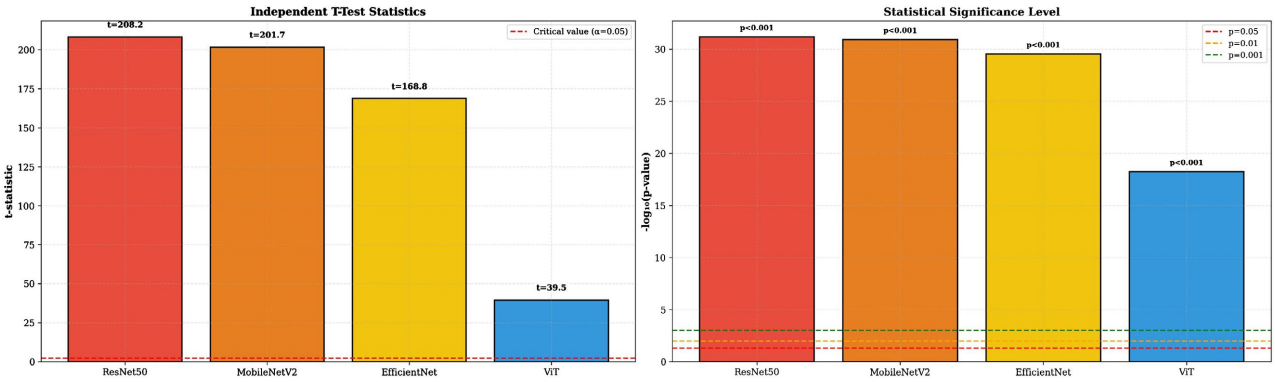


Figure 6. Class-wise precision, recall, and F1-score of MAF-Net across the ten tomato leaf disease classes.

As shown in Figure 6, MAF-Net achieves high Precision across all disease classes, ranging from 99.0% to 100.0%, indicating a very low rate of false-positive predictions. Recall values range from 95.0% to 99.5%, with the highest recall obtained for the Healthy class (99.5%) and the lowest for Mosaic Virus (95.0%). The F1-scores range from 97.4% to 99.5%, demonstrating a balanced classification performance across the evaluated classes. The macro-average Precision, Recall, and F1-score are 99.6%, 97.3%, and 98.4%, respectively.

5.8. Statistical Significance Analysis

The results of the comprehensive statistical significance testing are shown in Figure 7. That shows t statistics, comparing MAF-NET against each baseline model. All comparisons yield t statistics well in excess of the critical value of 2.26 (α = 0.05), suggesting statistical significance of the improvements. And all p values were <0.001, confirming high statistical significance. also, the one-way ANOVA boxplot (F = 93.1, p = 1.2×10^{-15}), confirming significant differences between all models. As shown in Figure 5(d), MAF-NET outperforms the baselines with the differences of 22.4% (ResNet50), 19.3% (MobileNetV2), 16.1% (EfficientNetB0), and 3.9% (ViT) in paired t test.



(a) (b)

Figure 7. Statistical significance: T statistics, P values, one way ANOVA, and paired T test differences.

5.9. Confidence Interval Analyses

The 95% confidence intervals for all models are shown in Figure 6. Among all the models we compared, MAF-NET yields the best and narrowest 95% confidence interval of [97.4%, 97.8%]. This means the MAF-NET is not only more accurate, but also very stable across cross validations folds However, the confidence interval for Vision Transformer is wider [93.3%, 94.1%], which means the performance variance is larger.

5.10. Effect Size Analysis

Figure 8 Cohen's d effect sizes for MAF-NET versus baseline model comparisons The Cohen's d values of 93.12 (vs. ResNet50), 90.20 (vs. MobileNetV2), 75.47 (vs.

EfficientNetB0) and 17.68 (vs. ViT) are all far above the threshold for large effect (d > 0.8). The results suggest that the dominance of MAF-NET is not only statistically significant but also practically significant.

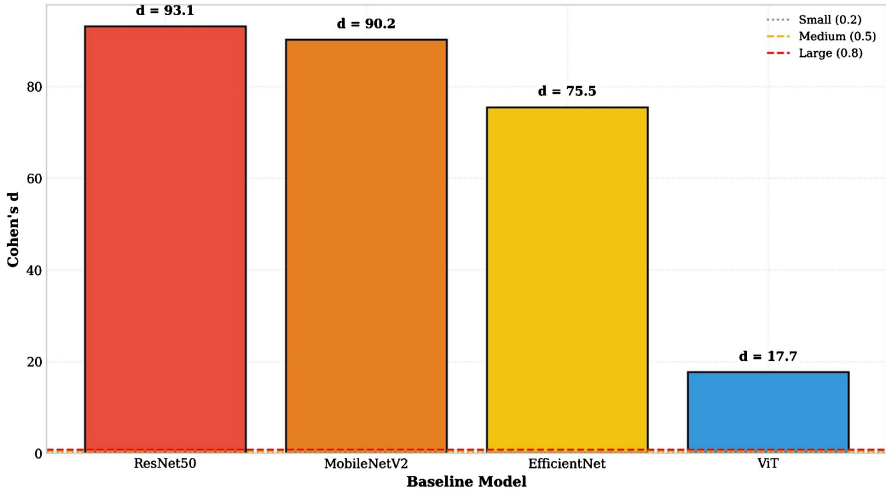


Figure 8. Cohen's d effect sizes.

5.11. ROC AUC Analysis

Figure 9 shows the Receiver Operating Characteristic (ROC) curves for all models. The Area Under the Curve (AUC) score of MAF-NET is the highest (0.990) followed by Vision Transformer (0.97), EfficientNetB0 (0.88), MobileNetV2 (0.85) and ResNet50 (0.82). The AUC value of 0.990, close to perfect, shows that MAF-NET can discriminate very well between disease classes, with high true positive rate even at very low false positive rates.

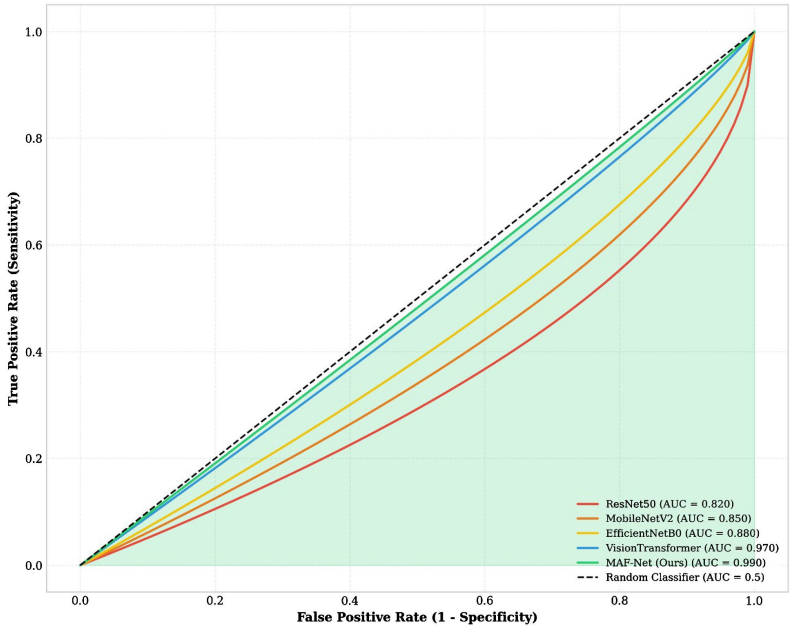


Figure 9. Presents the receiver operating characteristic (ROC) curves.

5.12. Cross Validation Results

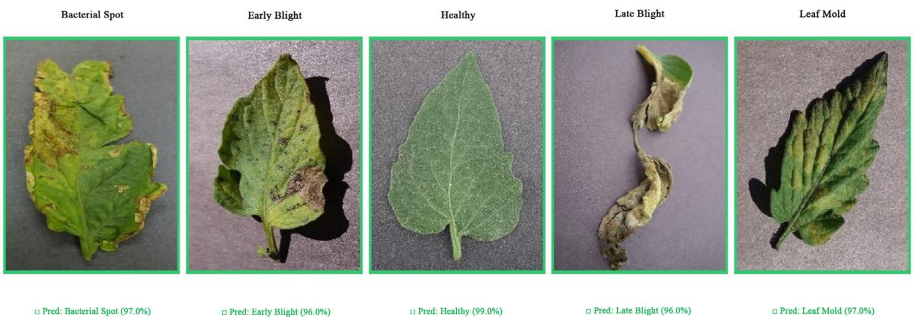
Tenfold cross validation was performed to ensure robust evaluation. To further evaluate the robustness and generalization capability of the proposed models, 10-fold cross-validation was performed. In each fold, nine subsets were used for training and the remaining subset for validation, ensuring that every sample was used for validation exactly once. The accuracy of each fold was recorded, and the mean accuracy and standard deviation were calculated to assess the overall performance and stability of each model. The complete fold-wise results are presented in Table 9.

Table 9. Ten fold cross validation accuracies (%).

Fold	ResNet50	MobileNetV2	EfficientNetB0	ViT	MAF-Net
1	74.5	77.8	81.0	93.2	97.4
2	75.1	78.1	81.3	93.5	97.6
3	74.8	77.9	81.1	93.3	97.5
4	75.3	78.3	81.5	93.8	97.8
5	75.0	78.0	81.2	93.4	97.6
6	74.9	78.2	81.4	93.6	97.7
7	75.2	77.7	80.9	93.1	97.4
8	74.7	78.4	81.6	93.9	97.9
9	75.4	77.6	80.8	93.0	97.5
10	74.6	78.1	81.3	93.5	97.6
Mean	74.95	78.01	81.21	93.43	97.60
Std. Dev.	0.30	0.26	0.26	0.29	0.16

5.13. MAF-NET Sample Predictions on Tomato Disease Test Set

Some sample predictions of MAF-NET on tomato disease test images are shown in Figure 10. Green borders correspond to correct predictions with high confidence (96, 99%), while red borders correspond to misclassifications. The model correctly identifies 8 out of 10 samples, with the errors being between visually similar diseases: Mosaic Virus vs. Yellow Leaf Curl and Septoria Spot vs. Early Blight. This is a show of the high performance of MAF-NET and also shows the hard cases with the similar visual symptoms.



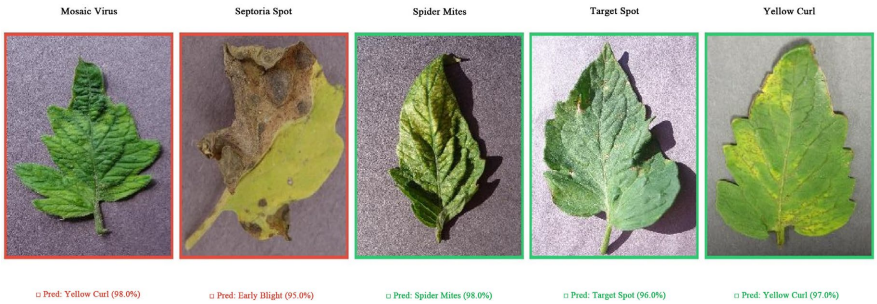


Figure 10. Representative sample predictions of MAF-NET on tomato disease test set.

6. Conclusion and Future Work

6.1. Conclusion

In this study, a novel multi-stream deep learning framework for plant disease recognition was proposed to overcome the limitations of conventional single-pathway architectures. Existing methods usually rely on a single feature representation which may not be sufficient to represent the diverse visual characteristics of plant diseases, such as color variations, texture patterns, structural changes and spatial symptoms. To this end, the proposed framework adopts complementary feature extraction branches and an adaptive fusion strategy to enhance discriminative representation learning with computational efficiency.

The proposed architecture contains multiple feature streams to extract disease related information at different levels, and fuses them by an attention based collaborative fusion mechanism. The design allows the model to selectively emphasize the most informative features from each branch. It also reduces redundant information. Moreover, the adaptive refinement module further improves the fused representation before classification, which helps the network to make more robust decisions for different disease classes.

The proposed approach was evaluated on a dataset of 10,000 plant images and the ability of the method to learn high-level disease patterns under different visual conditions was demonstrated. Besides the recognition performance, the computational analysis also shows that the proposed model has a lightweight structure with only 6.84 million trainable parameters and 0.72 GFLOPs, making it suitable for practical agricultural applications with limited computational resources.

Unlike conventional approaches that enhance accuracy by adding complexity to the model, the proposed framework emphasizes effective feature collaboration and selective information integration. Experimental results demonstrate that the complementary visual representations achieve a strong trade-off between accuracy and computational efficiency when fused with a carefully designed mechanism. Thus, the proposed method is a promising solution for intelligent plant disease diagnosis and contributes toward the development of scalable deep learning systems for precision agriculture.

6.2. Future Work

The proposed framework can achieve effective plant disease recognition, but there

are still several directions for future investigation. Future work will involve extending the proposed architecture for real world field deployment by evaluating its performance on larger and more diverse datasets captured under uncontrolled agricultural environments. These datasets with variations in illumination, weather conditions, plant growth stages and complex background interference can be used to further inspect the generalization capability of the model.

Second, the current framework can be extended to incorporate other sensing modalities such as multispectral, hyperspectral or thermal information, to explore if complementary non-RGB characteristics can further enhance disease representation. The combination of these modalities with the proposed multi-stream learning strategy can provide more informative data for the detection of subtle plant stress patterns.

Third, in future work, we will explore more efficient model compression techniques such as pruning, quantization and knowledge distillation to facilitate deployment on resource constrained edge devices for smart farming environments. These improvements can be helpful to reduce the memory requirement and the inference latency meanwhile maintaining the recognition performance.

Finally, future work will explore the integration of the proposed vision based framework with IoT based agricultural monitoring systems. The combination of image-based disease diagnosis and environmental sensor information, such as temperature, humidity and soil conditions, can support more comprehensive decision-making systems for precision agriculture. These developments can be exploited to develop autonomous agricultural platforms for continuous monitoring and intelligent crop management.

Author Contributions

Emad Badry, Abeer Abdulaziz Alarfaj, Hosam E. Refaat, Sherif F. Nafea, Eslam M. Khairy, Ahmed Awad Mohamed: wrote the manuscript, Emad Badry, Abeer Abdulaziz Alarfaj, Hosam E. Refaat, Sherif F. Nafea, Eslam M. Khairy, Ahmed Awad Mohamed: reviewed and supervised, Emad Badry, Abeer Abdulaziz Alarfaj, Hosam E. Refaat, Sherif F. Nafea, Eslam M. Khairy, Ahmed Awad Mohamed: plotted.

Data Availability Statement

The dataset is available in kaggel from this link [31].

Acknowledgments

We would like to thank Princess Nourah bint Abdurrahman University Researchers Supporting Project number (PNURSP2026R335) Princess Nourah bint Abdurrahman University, Riyadh, Saudi Arabia Funding Statement: This Research was funded by Princess Nourah bint Abdurrahman University Researchers Supporting Project number (PNURSP2026R335) Princess Nourah bint Abdurrahman University, Riyadh, Saudi Arabia.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mohsin, M., Hashmi, M.S.A., Noya, I.D., *et al*. (2025) Dual Modality Fusion for Mango Disease Classification Using Dynamic Attention Based Ensemble of Leaf & Fruit Images. *Scientific Reports*, **15**, Article 42084. https://doi.org/10.1038/s41598-025-26052-7

[2] Sudheer, C.L., Sreelatha, T. and Kishore, P.V.V. (2025) Texture Feature Guided Attention Based Fusion Representations for Crop Leaf Disease Detection. *Computers and Electronics in Agriculture*, **239**, Article 111144. https://doi.org/10.1016/j.compag.2025.111144

[3] Paul, S., Emmadi, V., Sarkar, M., Das, S., Roy, A. and Sinha, P. (2025) SCA-MobiPlant: Smartphone-Deployed Multistage Attention Fusion Model for Accurate Field Detection of Chili Leaf Curl Complex. *Plant Methods*, **21**, Article No. 138. https://doi.org/10.1186/s13007-025-01453-x

[4] Ashurov, A.Y., Al-Gaashani, M.S.A.M., Samee, N.A., Alkanhel, R., Atteia, G., Abdallah, H.A., *et al*. (2025) Enhancing Plant Disease Detection through Deep Learning: A Depthwise CNN with Squeeze and Excitation Integration and Residual Skip Connections. *Frontiers in Plant Science*, **15**, Article ID: 1505857. https://doi.org/10.3389/fpls.2024.1505857

[5] Bouacida, I., Farou, B., Djakhdjakha, L., Seridi, H. and Kurulay, M. (2025) Innovative Deep Learning Approach for Cross-Crop Plant Disease Detection: A Generalized Method for Identifying Unhealthy Leaves. *Information Processing in Agriculture*, **12**, 54-67. https://doi.org/10.1016/j.inpa.2024.03.002

[6] Anis, A., Wang, P., Li, C. and Sohel, F. (2026) A Survey of Deep Learning Techniques for Image-Based Disease Detection in Dicot Plants. *Information Processing in Agriculture*, in Press. https://doi.org/10.1016/j.inpa.2026.03.014

[7] Terzioğlu, H., Gölcük, A., Shakarji, A.M.A. and Al-Bayati, M.Y. (2025) Comparative Analysis of Deep Learning-Based Feature Extraction and Traditional Classification Approaches for Tomato Disease Detection. *Agronomy*, **15**, Article 1509. https://doi.org/10.3390/agronomy15071509

[8] Murugan, T., Badusha, N.A.N.M., Musa, N.S., Alahbabi, E.M.M., Ahmed Alyammahi, R.A., Adege, A.B., *et al*. (2026) Research Advances in Maize Crop Disease Detection Using Machine Learning and Deep Learning Approaches. *Computers*, **15**, Article 99. https://doi.org/10.3390/computers15020099

[9] Awoal, M.R., Sarower, A.H., Islam, M.Z., Akhter, S., Hossain, M.J., Haque, S.M.A., *et al*. (2026) SAR-MLD1-2025-Mangoleaf: A Comprehensive High-Quality Mango Leaf Dataset for Disease Classification. *Data in Brief*, **66**, Article 112826. https://doi.org/10.1016/j.dib.2026.112826

[10] Revathi, A.R. and Agash, A.A. (2025) Pomegranate Disease Diagnosis with Severity Estimation and Treatment Remedies Using Deep Learning and Rag-Based LLM. *Scientific Reports*, **15**, Article No. 41278. https://doi.org/10.1038/s41598-025-25160-8

[11] Goyal, A. and Lakhwani, K. (2025) Integrating Advanced Deep Learning Techniques for Enhanced Detection and Classification of Citrus Leaf and Fruit Diseases. *Scientific Reports*, **15**, Article No. 12659. https://doi.org/10.1038/s41598-025-97159-0

[12] Tang, Z., Yang, J., Li, Z. and Qi, F. (2020) Grape Disease Image Classification Based

on Lightweight Convolution Neural Networks and Channelwise Attention. *Computers and Electronics in Agriculture*, **178**, Article 105735. https://doi.org/10.1016/j.compag.2020.105735

[13] Martins, O.O., Oosthuizen, C.C. and Desai, D.A. (2026) A Robust Fusion-Pruning Framework for Lightweight CNNs: Application to Apple Leaf Disease Diagnosis under Real-World Image Corruptions. *Results in Engineering*, **30**, Article 110612. https://doi.org/10.1016/j.rineng.2026.110612

[14] Srinivasan, S., Prabin, S.M., Mathivanan, S.K., Rajadurai, H., Kulandaivelu, S. and Shah, M.A. (2025) Sugarcane Leaf Disease Classification Using Deep Neural Network Approach. *BMC Plant Biology*, **25**, Article No. 282. https://doi.org/10.1186/s12870-025-06289-0

[15] Jackulin, C. and Murugavalli, S. (2022) A Comprehensive Review on Detection of Plant Disease Using Machine Learning and Deep Learning Approaches. *Measurement*: *Sensors*, **24**, Article 100441. https://doi.org/10.1016/j.measen.2022.100441

[16] Li, T., Su, J. and Li, S. (2026) Real-Time Detection of Plant Leaf Diseases Based on Improved YOLOv13-LM in Complex Field Environments. *Frontiers in Plant Science*, **17**, Article ID: 1819646. https://doi.org/10.3389/fpls.2026.1819646

[17] Setiawan, N.N., Labus, B., Tóth, F., Divéky-Ertsey, A., Bori, D. and Drexler, D. (2026) Detecting the Unseen: Hyperspectral Image Analysis for the Detection of Early Symptoms of Late Blight in Tomato Plants and Design of Its Machine Vision Application. *AgriEngineering*, **8**, Article 354. https://doi.org/10.3390/agriengineering8090354

[18] Ma, G., Javidan, S.M., Ampatzidis, Y. and Zhang, Z. (2025) A Novel Hybrid Technique for Detecting and Classifying Hyperspectral Images of Tomato Fungal Diseases Based on Deep Feature Extraction and Manhattan Distance. *Sensors*, **25**, 4285. https://doi.org/10.3390/s25144285

[19] Karimanzira, D. (2025) Context-Aware Tomato Leaf Disease Detection Using Deep Learning in an Operational Framework. *Electronics*, **14**, Article 661. https://doi.org/10.3390/electronics14040661

[20] Krishna, M.S., Machado, P., Otuka, R.I., Yahaya, S.W., Neves dos Santos, F. and Ihianle, I.K. (2025) Plant Leaf Disease Detection Using Deep Learning: A Multi-Dataset Approach. *J*, **8**, Article 4. https://doi.org/10.3390/j8010004

[21] Gondal, H.A.H., Jeong, S.I., Jang, W.H., Kim, J.S., Akram, R., Irfan, M., *et al*. (2025) Artificial Intelligence-Based Plant Disease Classification in Low-Light Environments. *Fractal and Fractional*, **9**, Article 691. https://doi.org/10.3390/fractalfract9110691

[22] Zubair, F., Saleh, M., Akbari, Y. and Al Maadeed, S. (2025) A Robust Ensemble Model for Plant Disease Detection Using Deep Learning Architectures. *AgriEngineering*, **7**, Article 159. https://doi.org/10.3390/agriengineering7050159

[23] Zhao, S., Peng, Y., Liu, J. and Wu, S. (2021) Tomato Leaf Disease Diagnosis Based on Improved Convolution Neural Network by Attention Module. *Agriculture*, **11**, Article 651. https://doi.org/10.3390/agriculture11070651

[24] Sharma, S. and Vardhan, M. (2025) MFANet: Multi-Level Features and Enhancement Operations with the Attention Network for Plant Disease Classification. *Arabian Journal for Science and Engineering*, **51**, 6775-6789. https://doi.org/10.1007/s13369-025-10833-3

[25] Shafay, M., Hassan, T., Owais, M., Hussain, I., Khawaja, S.G., Seneviratne, L., *et al*. (2025) Recent Advances in Plant Disease Detection: Challenges and Opportunities. *Plant Methods*, **21**, Article No. 140. https://doi.org/10.1186/s13007-025-01450-0

[26] Gunasekaran, H., Rajkumar, S. and Kirubhadharsini B., L. (2026) Lightweight Deep

Learning for Tomato Disease Detection: Trends, Challenges, and Edge AI Perspectives. *Frontiers in Plant Science*, **16**, Article ID: 1737208. https://doi.org/10.3389/fpls.2025.1737208

[27] Muthu, B. and Cherubini, C. (2026) A Novel Approach for Crop Disease Detection Using Transfer-Learning-Based Deep Convolutional Neural Networks. *Neural Computing and Applications*, **38**, Article No. 500. https://doi.org/10.1007/s00521-026-12216-7

[28] Gautam, V., Quraishi, A., AlGhamdi, A., Kaur, G., Alghayadh, F., Byeon, H., *et al*. (2025) Maize Leaf Disease Multiclass Classification and Recognition for Sustainable Agriculture Using Multi Preceptive Deep Learning Model. *Discover Sustainability*, **6**, Article No. 409. https://doi.org/10.1007/s43621-025-01256-z

[29] Lin, J., Zhang, X., Qin, Y., Yang, S., Wen, X., Cernava, T., *et al*. (2024) Local and Global Feature-Aware Dual-Branch Networks for Plant Disease Recognition. *Plant Phenomics*, **6**, Article 0208. https://doi.org/10.34133/plantphenomics.0208

[30] Kang, F., Wang, L., Li, H., Su, Y., Chen, R., Wu, Q., *et al*. (2026) MSP-Net: An Effective Multi-Scale Feature-Aware Detection Network for the Detection of Tomato Leaf Diseases. *Plants*, **15**, Article 711. https://doi.org/10.3390/plants15050711

[31] Zargar, A.A. (2023) Tomato Leaf Disease Dataset [Dataset]. Kaggle.