Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 August 2025
DOI 10.3389/fpls.2025.1638520

OPEN ACCESS

EDITED BY
Vinayakumar Ravi,
Prince Mohammad bin Fahd University,
Saudi Arabia

REVIEWED BY
Bahar Mahmud,
Western Michigan University, United States
Vasudha Vedula,
University of Texas of the Permian Basin,
United States
Gowtham M.,
JSS Science and Technology University, India

*CORRESPONDENCE
Ebru Ergün
ebru.yavuz@erdogan.edu.tr

RECEIVED 04 June 2025
ACCEPTED 11 July 2025
PUBLISHED 01 August 2025

CITATION
Ergün E (2025) Attention-enhanced hybrid deep learning model for robust mango leaf disease classification via ConvNeXt and vision transformer fusion.
Front. Plant Sci. 16:1638520.
doi: 10.3389/fpls.2025.1638520

Attention-enhanced hybrid deep learning model for robust mango leaf disease classification via ConvNeXt and vision transformer fusion

Ebru Ergün*

Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Recep Tayyip Erdogan University, Rize, Türkiye

Mango is a crop of vital agronomic and commercial importance, particularly in tropical and subtropical regions. Accurate and timely identification of foliar diseases is essential for maintaining plant health and ensuring sustainable agricultural productivity. This study proposes MangoLeafCMDF-FAMNet (cross-modal dynamic fusion with feature attention module (FAM) network), an advanced, hybrid, deep-learning framework designed for the multi-class classification of mango leaf diseases. The model combines two state-of-the-art feature extractors, ConvNeXt and Vision Transformer, to capture local fine-grained textures and global contextual semantics simultaneously. To further improve feature discrimination, a FAM inspired by squeeze-and-excitation networks is integrated into each stage of the backbone. This module adaptively recalibrates channel-wise feature responses to highlight disease-relevant cues while suppressing irrelevant background noise. A novel cross-modal dynamic fusion strategy unifies the complementary strengths of both branches, resulting in highly robust and discriminative feature embeddings. The proposed model was rigorously evaluated using comprehensive metrics such as classification accuracy (CA), recall, precision, Matthews correlation coefficient (MCC) and Cohen's kappa score on three benchmark datasets: MangoLeafDataset1 (8 classes), MangoLeafDataset2 (5 classes) and MangoLeafDataset3 (8 classes). The experimental results consistently demonstrate the superiority of MangoLeafCMDF-FAMNet over the existing baseline models. It achieves exceptional CA values of 0.9978, 0.9988 and 0.9943 across the respective datasets, alongside strong MCC and Cohen's kappa scores. These results highlight the effectiveness and generalizability of the proposed framework for automated mango leaf disease diagnosis and contribute to advancing deep learning applications in precision plant pathology.

KEYWORDS

agricultural imaging, ConvNeXt, cross-modal dynamic fusion, disease classification, mango leaf, vision transformer

1 Introduction

The mango is of great agronomic and economic importance, particularly in tropical and subtropical regions, where it is one of the most widely cultivated fruit crops (Zhang et al., 2025). However, mango plants' productivity and health are persistently threatened by foliar diseases, which impair photosynthetic efficiency and lead to significant reductions in yield and fruit quality. Although traditional approaches to disease diagnosis are still widely used, they often involve subjective assessments, delayed response times and a heavy dependence on expert knowledge. These limitations highlight the urgent need for automated, accurate and scalable diagnostic tools to support the timely and objective management of mango diseases.

In recent years, deep learning (DL) techniques have transformed plant disease detection by enabling complex, hierarchical patterns to be extracted directly from image data. Unlike conventional handcrafted approaches, DL models have demonstrated superior performance in various agricultural vision tasks. However, classifying mango leaf diseases is challenging due to their intricate visual symptoms, similarities between classes and variations within classes that frequently occur across disease types (Shehu et al., 2025; Wei et al., 2025). These complexities necessitate the development of more advanced architectures that can effectively learn both fine-grained local textures and high-level semantic features.

In response to these challenges, we propose MangoLeafCMDF-FAMNet (cross-modal dynamic fusion (CMDF) with feature attention module (FAM) network), a novel hybrid DL framework specifically designed for multi-class mango leaf disease classification. This architecture integrates ConvNeXt and Vision Transformer (ViT) as dual feature extractors, combining the strengths of convolutional inductive biases and transformer-based global attention mechanisms. The proposed model uses a CMDF strategy to combine texture- and semantic-level information into a coherent, enriched feature representation space. To further enhance feature expressiveness, a FAM, inspired by squeeze-and-excitation (SE) networks, is incorporated at each stage. This module adaptively recalibrates channel-wise feature responses to prioritize disease-relevant patterns while suppressing irrelevant background noise.

To comprehensively evaluate the performance of MangoLeafCMDF-FAMNet, we conduct extensive experiments on three publicly available mango leaf disease datasets—MangoLeafDataset1 (8 classes), MangoLeafDataset2 (5 classes), and MangoLeafDataset3 (8 classes). The model's effectiveness is quantified using multiple evaluation metrics, including classification accuracy (CA), recall (RCL), precision (PRC), Matthews correlation coefficient (MCC), and Cohen's kappa score. The experimental results consistently demonstrate that the proposed method significantly outperforms conventional baseline models across all datasets, achieving high CA and strong correlation measures.

The main contributions of this work are as follows:

- We introduce MangoLeafCMDF-FAMNet, a novel hybrid DL architecture that combines ConvNeXt and ViT with FAM for enhanced hierarchical feature representation.
- We design a CMDF strategy that effectively fuses local texture information with global contextual features, leading to more robust representations.
- We perform a thorough evaluation across multiple public datasets, establishing the superior classification performance of the proposed model in terms of CA, RCL, PRC, MCC, and kappa.
- We offer a generalizable and scalable framework with practical implications for automated mango disease diagnosis, and the potential for adaptation to other plant disease classification tasks.

The remainder of this paper is structured as follows: Section 2 provides a thorough review of existing research for plant disease detection. Section 3 outlines the materials and methods employed in this study, providing detailed descriptions of the dataset and the proposed hybrid deep learning and feature selection framework. Section 4 reports the experimental results, alongside performance evaluation metrics and comparative analyses. Section 5 concludes the paper by summarizing the main findings and highlighting potential future research directions. Finally, Section 6 critically discusses the study's limitations and underlying assumptions, as well as its practical implications for real-world agricultural applications.

2 Review of existing approaches

Recent advances in computer vision and DL have significantly accelerated the development of automated tools for plant disease diagnosis. convolutional neural networks (CNNs) and transformer-based models, especially ViTs, have emerged as dominant paradigms in plant pathology research due to their ability to extract discriminative spatial and semantic patterns from complex visual data (Chen et al., 2024). However, existing studies on mango leaf disease classification have faced multiple challenges that limit their practical utility and generalizability. Among these studies, Rao et al. (2021) initiated one such effort by leveraging the PlantVillage dataset, applying the AlexNet architecture, and achieving classification accuracies of 0.9900 for grape leaves and 0.8900 for mango leaves. Building on this, Arivazhagan and Ligi (2018) utilized a CNN-based approach on a six-class mango dataset and attained a commendable accuracy of 0.9667. In contrast, Mia et al. (2020) combined artificial neural networks and support vector machines, achieving 0.8000 accuracy in detecting four disease classes and healthy leaves. A more advanced ensemble strategy was introduced by Gautam et al. (2024), who developed a Stacked Ensemble Deep Neural Network that integrated multiple DNNs with classical ML classifiers, yielding a high accuracy of 0.9857 across eight disease classes. Similarly, Saleem et al. (2021a) investigated disease detection using canonical correlation analysis (CCA)-based feature fusion and found cubic SVM to deliver the highest performance. In a follow-up study, Saleem et al. (2021b) further proposed the FrCNet model for lesion segmentation and, after combining it with CCA feature fusion and classification via quadratic and cubic SVMs, achieved 0.9890 accuracy for binary disease-versus-healthy discrimination. Continuing the exploration of CNN variants, Varma et al. (2025) benchmarked several pretrained architectures, with InceptionV3 achieving the highest accuracy at 0.9987. Meanwhile, Patel et al. (2024) introduced a hybrid framework combining Total Variation Filter-based variational mode decomposition with DenseNet121 and VGG-19, achieving 0.9885 CA. This fusion approach notably improved feature interpretability and robustness against noise. Hossain et al. (2024) evaluated ViTs against well-established CNNs and proposed an optimized DeiT-based model, which outperformed all compared methods with a CA of 0.9975. Similarly, Mahmud et al. (2024) proposed DenseNet78, a lightweight variant of DenseNet tailored for mango leaf disease classification, reporting accuracies of 0.9947 for healthy and 0.9944 for diseased leaves. In practical implementations, Puranik et al. (2024) retrained MobileNetV3 on the MangoLeafBD dataset and embedded it within a mobile application, reaching 0.9800 accuracy and enabling real-time field diagnosis. Singh et al. (2024) adopted a transfer learning approach and proposed the DTLD model, which demonstrated strong multi-class classification performance with a peak accuracy of 0.9976 on a 4000-image dataset. Expanding on comparative model analysis, Bairwa et al. (2024) assessed multiple deep networks, finding ResNet50 to deliver the highest accuracy at 0.9912. Complementarily, Pratap and Kumar (2024) designed a CNN-based system incorporating transfer learning from VGG-16, GoogLeNet, MobileNet, YOLOv8, and EfficientNet, enabling effective classification of several mango diseases including Anthracnose, Gall Midge, and Powdery Mildew. Finally, Pahati et al. (2025) trained a Google Teachable Machine model on 4000 annotated images, obtaining an accuracy of 0.9960 and demonstrating high potential for democratized, user-friendly disease recognition platforms.

Most conventional CNN-based models, although capable of capturing local textures, fall short in modeling long-range dependencies—a critical requirement for accurately distinguishing visually similar diseases with subtle morphological variations. Transformer-based methods, while excellent at global context modeling, often lack the inductive biases necessary for fine-grained feature localization. As such, stand-alone CNN or ViT models struggle to deliver optimal performance across varying environmental conditions and disease stages observed in agricultural settings. For example, the study by Alamri et al. (2025) proposed a dual-branch architecture combining ConvNeXt and ViT to detect mango leaf and fruit diseases separately using the MangoLeafBD and SenMangoFruitDDS datasets. Their model achieved promising accuracy levels of 99.87% and 98.40% respectively, demonstrating the value of hybrid architectures in plant disease classification. However, their method did not incorporate any explicit attention mechanism to recalibrate the feature importance across network layers. Furthermore, their architecture processed the outputs of ConvNeXt and ViT using a static fusion approach, which may limit the adaptability of feature interactions during training.

By contrast, our proposed MangoLeafCMDF-FAMNet framework introduces several significant improvements to the original design. Firstly, inspired by SE networks, we incorporated a FAM at each stage to dynamically recalibrate channel-wise features. This enables the model to selectively emphasize disease-relevant information and suppress background noise. Secondly, instead of using a static feature aggregation strategy, our model uses a CMDF mechanism to adaptively combine spatial and semantic cues extracted from ConvNeXt and ViT backbones. This significantly improves the model's representational richness and robustness.

Moreover, MangoLeafCMDF-FAMNet was rigorously evaluated on three distinct datasets encompassing both 5-class and 8-class classification tasks. Experimental results demonstrated that our model consistently outperforms traditional CNNs, ViTs, and hybrid baselines—including the model by Alamri et al. (2025) —not only in terms of CA but also across comprehensive evaluation metrics such as MCC and kappa. The superior performance of our model, particularly under multi-class, real-world conditions, underscores its potential as a scalable and generalizable solution for precision agriculture.

Importantly, foliar disease diagnosis remains a crucial but underexplored area in the literature, especially concerning tropical crops such as mango. Leaf diseases are often early indicators of plant stress and can significantly affect fruit development and overall yield. Therefore, developing robust, accurate, and field-deployable diagnostic systems for leaf disease identification is critical for achieving sustainable agricultural outcomes. Our contribution lies not only in achieving state-of-the-art performance but also in offering a practical architecture that balances accuracy, computational efficiency, and adaptability, setting a new benchmark in artificial intelligence (AI)-assisted mango disease diagnosis.

3 Materials and methods

3.1 Description of dataset

3.1.1 MangoLeafDataset1

The Mango MLD dataset, (MLD1) curated by Shakib et al., served as one of the primary data sources in this study (Shakib et al., 2024). This publicly available dataset was meticulously compiled through an extensive field data acquisition campaign conducted across diverse mango orchards situated in Kushtia and Dhaka, Bangladesh. The primary objective of this collection effort was to capture high-quality, representative images of both healthy and diseased mango leaves under realistic agricultural conditions, thereby ensuring the ecological validity and practical relevance of the dataset for real-world disease classification tasks.

A total of 6,400 images were included in the dataset, uniformly distributed across eight diagnostic categories. Seven of these classes correspond to prevalent mango leaf diseases—Anthracnose, Bacterial Canker, Cutting Weevil, Die Back, Gall Midge, Powdery Mildew, and Sooty Mould—while the eighth class represents healthy leaves. To mitigate class imbalance and enable unbiased model training, each category contains exactly 800 images, making this dataset structurally balanced. The images were originally captured using an iPhone SE device at a native resolution of 3024 × 4032 pixels and subsequently downscaled to 240 × 240 pixels in

JPEG format. This resizing operation was performed to reduce memory overhead without significantly compromising visual quality or diagnostic features. Crucially, no synthetic augmentation was applied to the original images, preserving the integrity and authenticity of real-world leaf textures, color gradients, and lesion morphologies.

Figure 1 presents representative samples from each disease class, providing visual insight into the morphological and pathological variations captured in the dataset. Meanwhile, the corresponding distribution of class frequencies is detailed in Table 1, where the uniformity of sample counts across categories is explicitly demonstrated.

3.1.2 MangoLeafDataset2

As a complementary data source, the MangoLeafDataset2 (MLD2)—compiled and published by Nirob et al.—was incorporated into this study to further strengthen the reliability and generalizability of the proposed classification model (Nirob et al., 2024). This dataset offers a rich collection of high-resolution mango leaf images, originally captured between August 15 and August 29, 2023, in the mango cultivation fields of Supu Ashulia, Bangladesh. The data acquisition process was carried out under natural lighting and environmental conditions, ensuring that the captured leaf samples reflect real-world visual characteristics, including noise, background clutter, and variability in disease presentation.

The original dataset comprises 1,319 unique images, each with a standardized resolution of 1000 × 1000 pixels and stored in JPEG format. The dataset encompasses five key categories representing distinct pathological states of mango leaves: Anthracnose, Die Black, Gall Midge, Powdery Mildew, and Healthy. These categories were carefully selected based on the prevalence and diagnostic importance of the corresponding diseases in commercial mango production. To overcome the inherent class imbalance, present in the original dataset and to enhance the learning capability of deep models, a comprehensive data augmentation strategy was applied. Techniques such as horizontal and vertical flipping, arbitrary rotations, scaling, and mild intensity transformations were utilized to synthetically expand the dataset. As a result, each category was normalized to contain exactly 1,000 samples, thereby yielding a final augmented dataset comprising 5,000 images. Figure 2 illustrates representative image samples from each of the five classes, providing a visual overview of the phenotypic diversity embedded within the dataset. Table 2 summarizes the distribution of both original and augmented images per class.

3.1.3 MangoLeafDataset3

To further enhance the robustness and cross-dataset generalizability of the proposed classification framework, this study integrated the MangoLeafDataset3 (MLD3), meticulously curated by Rahman et al. and publicly released in November 2024 (Rahman et al., 2024). This dataset serves as a significant and diverse benchmark resource for intelligent agricultural analysis, particularly in the field of mango leaf disease recognition. The data acquisition phase was carried out over a period of 20 consecutive days, from October 15 to November 4, 2024, in two distinct agroecological regions of Bangladesh—Kashinathpur (Pabna) and Changao (Savar, Dhaka)—to capture a wide spectrum of environmental and disease conditions. The dataset is composed of two main subsets: 2,336 raw images captured under natural lighting conditions using mobile phone cameras, and 12,730 synthetically augmented images generated through comprehensive data enhancement techniques. These augmentations include but are

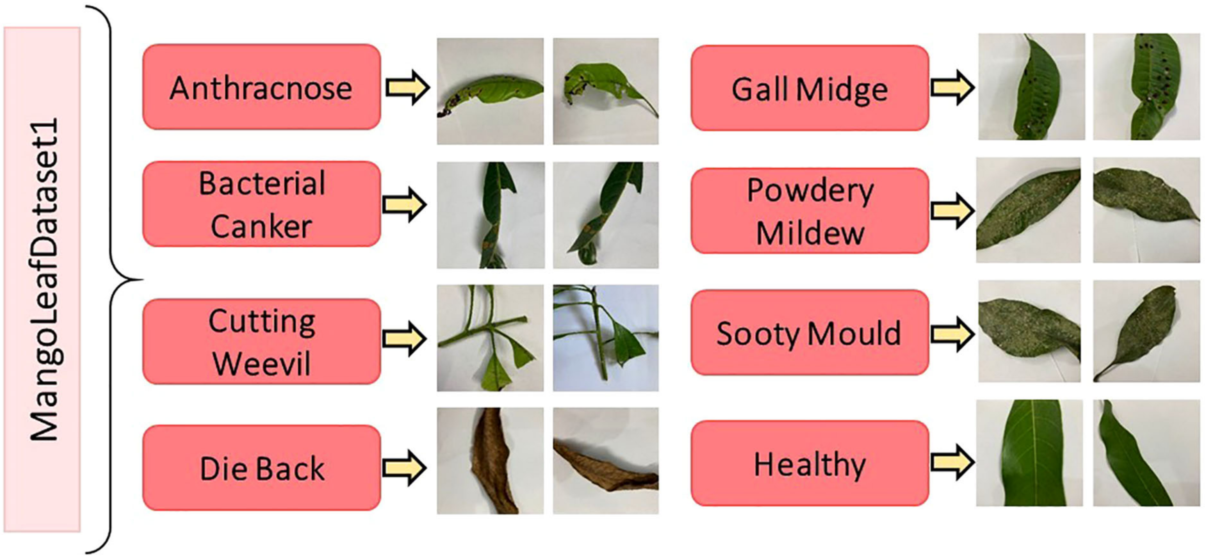


FIGURE 1
Representative images from the MLD1 dataset, illustrating both healthy mango leaves and leaves affected by seven distinct diseases (Shakib et al., 2024).

TABLE 1 Class-wise distribution of images in the MLD1 (Shakib et al., 2024).

Class Name	MLD1
	Number of images
Anthracnose	800
Bacterial Canker	800
Cutting Weevil	800
Die Back	800
Gall Midge	800
Powdery Mildew	800
Sooty Mould	800
Healthy	800
Total	6400

not limited to affine transformations, horizontal/vertical flips, minor brightness and contrast shifts, and random cropping, all designed to introduce variability and enrich the dataset's learning potential without compromising biological authenticity. All images are categorized into eight distinct classes, representing seven pathological categories—Anthracnose, Bacterial Canker, Cutting Weevil, Die Back, Gall Midge, Powdery Mildew, and Sooty Mould—along with one Healthy class.

The original class distribution, prior to augmentation, is intentionally preserved to reflect natural disease occurrence rates. However, the expanded dataset introduces balance and diversity necessary for training deep neural models effectively. A comprehensive breakdown of the image count per class is provided in Table 3, while Figure 3 visually showcases representative samples from each class, highlighting inter-class visual variability and intra-class complexity.

3.2 Research methodology framework

In this study, a novel hybrid DL architecture named MangoLeafCMDF-FAMNet was proposed to address the complex problem of mango leaf disease classification. The methodology capitalized on the complementary strengths of two advanced feature extractors—ConvNeXt and ViT—to capture fine-grained texture details as well as global contextual dependencies inherent in leaf imagery. The overall flow of the proposed method is illustrated in Figure 4. To train and validate the proposed framework, three publicly available datasets were employed: the 8-class MLD1, 5-class MLD2, and 8-class MLD3. All image samples were preprocessed with standard normalization and resized to a

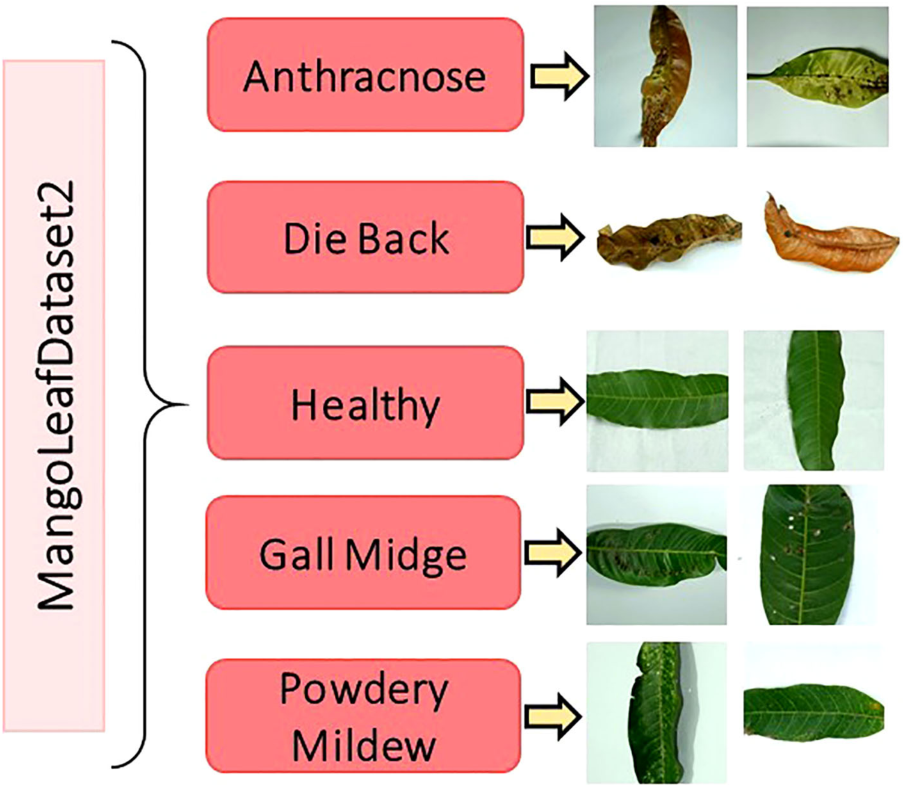


FIGURE 2
Selected image samples from the MLD2 dataset, reflecting variations in leaf color, texture, and shape across four disease classes and healthy leaves (Nirob et al., 2024).

TABLE 2 Augmented image counts for each class in the MLD2 (Nirob et al., 2024).

Class Name	MLD2
	Number of images
Anthracnose	1000
Die Back	1000
Gall Midge	1000
Powdery Mildew	1000
Healthy	1000
Total	5000

TABLE 3 Augmented image counts per class in the MangoLeafDataset3 (Rahman et al., 2024).

Class Name	MLD3
	Number of images
Anthracnose	1749
Bacterial Canker	2534
Cutting Weevil	1583
Die Back	1280
Gall Midge	2233
Powdery Mildew	776
Sooty Mould	1325
Healthy	1250
Total	12730

uniform resolution of 224 × 224 pixels to ensure consistency across training folds. Data augmentation techniques were deliberately excluded to assess the raw generalization power of the model.

The MangoLeafCMDF-FAMNet model integrated a ConvNeXt-Tiny backbone pre-trained on ImageNet as its local feature extractor. Its final classification head was removed and replaced with a global average pooling layer, followed by a custom FAM inspired by SE networks. In parallel, a lightweight ViT was employed to model long-range semantic interactions, with its outputs refined through a 1D feature-wise attention mechanism designed to amplify class-relevant representations. After extracting the deep features from both ConvNeXt and ViT branches, a CMDF strategy was applied. This strategy concatenated the learned embeddings and passed them through a projection layer, resulting in a unified 1024-dimensional representation. The fused vector was then forwarded to a fully connected classifier to generate final class predictions.

The performance evaluation of the proposed model was conducted using a stratified 5-fold cross-validation protocol (5-FCVP) to ensure reliable and unbiased assessment. In each fold, the dataset was partitioned into distinct training and validation subsets while preserving class distribution. During training, the model parameters were optimized using the AdamW optimizer, configured with a learning rate of 0.00005 and a weight decay coefficient of 0.0001 to promote generalization. The cross-entropy loss function served as the optimization objective, guiding the network's learning process. For each validation phase, a comprehensive set of evaluation metrics was computed, including CA, RCL, PRC, MCC, and kappa, to provide a multi-faceted performance analysis. Additionally, confusion matrices were generated for each fold to reveal class-specific prediction behaviors. To qualitatively investigate the separability of learned

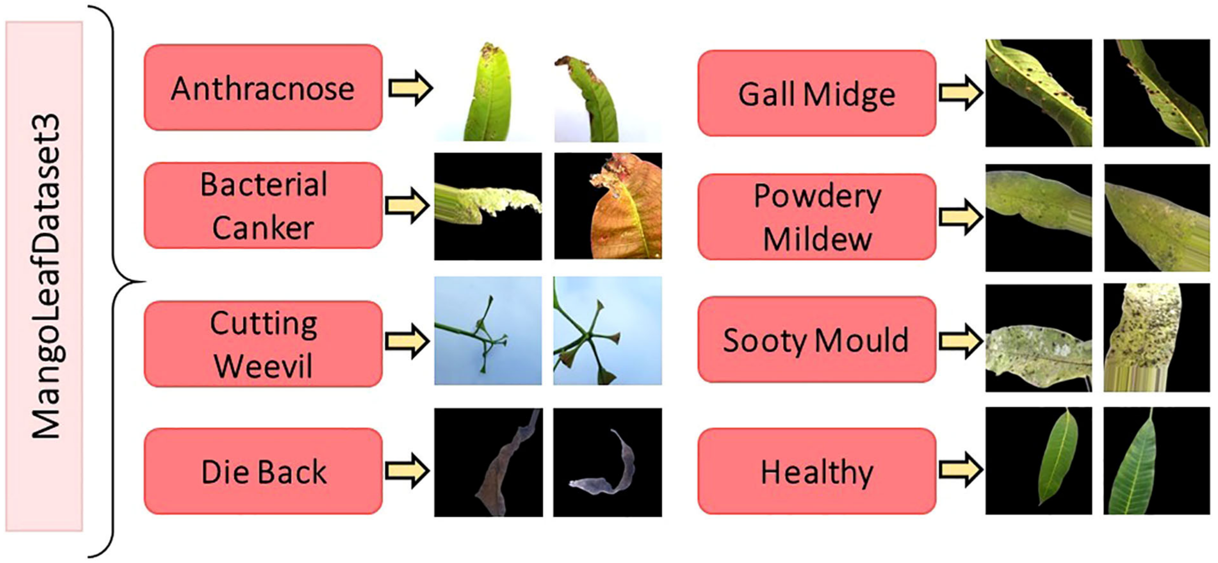


FIGURE 3
Image samples from the MLD3 dataset highlighting intra-class variability and visual diversity, which pose additional challenges for robust disease classification (Rahman et al., 2024).

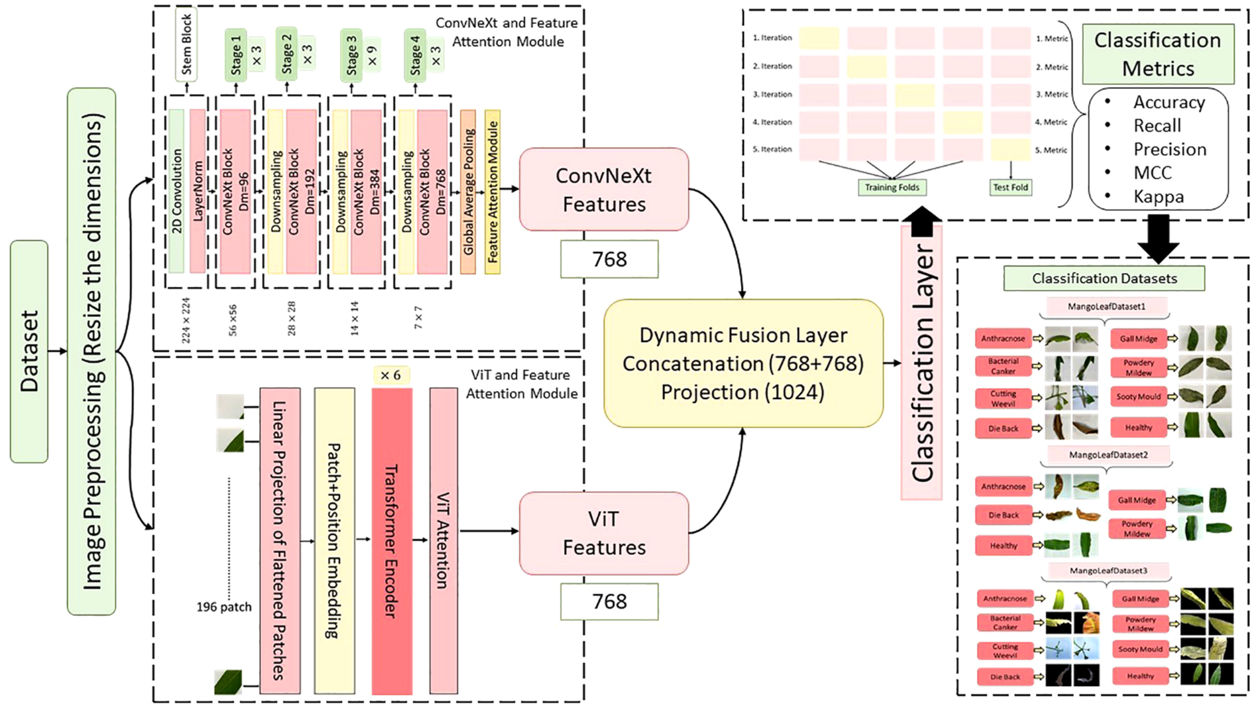


FIGURE 4
Overview of the proposed MangoLeafCMDF-FAMNet architecture, illustrating the dual-branch design consisting of ConvNeXt and ViT, integrated with FAM and a CMDF mechanism for feature integration and disease classification.

features, high-dimensional embeddings were projected into a two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE), offering visual insight into the model's discriminative capability.

3.3 ConvNeXt backbone architecture

In this study, ConvNeXt was selected as one of the core backbone networks of the proposed CMDF-Net due to its ability to effectively extract hierarchical features from input images by leveraging the design principles of both ResNet and transformer-based architectures. ConvNeXt is a convolutional neural network that modernizes the classic ResNet architecture through architectural refinements inspired by the success of ViTs, achieving a competitive balance between performance and efficiency in visual recognition tasks.

ConvNeXt comprises multiple stages, each containing a sequence of blocks designed to progressively capture low-level to high-level semantic features (Fu et al., 2025). Each block within ConvNeXt replaces the traditional bottleneck structure of ResNet with a streamlined stack of operations, composed of a depthwise convolution (DWConv), a layer normalization (LN), a pointwise convolution (1×1 Conv), and a GELU activation function. Mathematically, the core block of ConvNeXt can be formulated as follows. Let $x \in R^{H \times W \times C}$ represent the input tensor, where H, W, and C denote the height, width, and number of channels, respectively. The transformation $f(x)$ within a ConvNeXt block is defined as Equation 1 (Ford et al., 2025).

$$f(x) = W_2 GELU(LN(W_1 \cdot DWConv(x))) \tag{1}$$

where, DWConv represents the depthwise convolutional operation with a kernel size of 7×7, designed to capture spatial correlations within each channel independently. W_1 and W_2 denote pointwise (1×1) convolution weights that project the input and output feature spaces. LN is the layer normalization function, which stabilizes training and accelerates convergence. GELU stands for Gaussian error linear unit, providing smoother activation compared to ReLU.

In this implementation, ConvNeXt was configured with the "ConvNeXt-Tiny" variant to ensure a balanced trade-off between computational cost and feature extraction capability. The network was divided into four stages, where each stage contains multiple ConvNeXt blocks and concludes with a downsampling layer that reduces the spatial resolution while increasing the channel dimension (Lu et al., 2025). The channel dimensions across the stages were configured as $[96, 192, 384, 768]$, and the number of blocks per stage were $[3, 3, 9, 3]$, respectively. To further enhance the representational capacity of ConvNeXt, we integrated a FAM at the output of each stage. Inspired by the SE networks, this module adaptively recalibrates the feature maps along the channel dimension. The mechanism operates in three steps: squeeze, excitation, and reweighting. Let $U \in R^{H \times W \times C}$ in denote the output feature map of a ConvNeXt stage. The channel-wise global descriptor z $\in R^C$ is obtained via global average pooling given as Equation 2 (Tao et al., 2022).

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_{i,\ j,\ c} \quad (2)$$

Then, the excitation step applies a gating mechanism using two fully connected layers with non-linearity is shown in Equation 3.

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \quad (3)$$

where, $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$ are the weights of the fully connected layers, δ is the ReLU activation function, σ is the sigmoid activation function, r is the reduction ratio (set to 16 in this study) controlling the bottleneck. Finally, the recalibrated feature map $\hat{U}$ is obtained by channel-wise multiplication is given in Equation 4.

$$\hat{U}_c = s_c \cdot U_c \quad (4)$$

This attention mechanism allows the network to selectively emphasize informative features while suppressing less useful ones, thereby boosting the model's ability to focus on disease-related patterns in mango leaf images. The output feature maps from all ConvNeXt stages, enhanced by their respective attention modules, are then passed to the fusion layer as shown Figure 5.

3.4 Vision transformer backbone architecture

As a complementary backbone to ConvNeXt, the ViT was employed in CMDF-Net to exploit the global context modeling capabilities of self-attention mechanisms. ViT treats images as sequences of non-overlapping patches, analogous to tokens in natural language processing, and applies standard Transformer encoders to capture long-range dependencies and global feature representations, which are crucial for identifying disease patterns distributed across different regions of mango leaves (Ergün, 2025). The input image $x \in R^{H \times W \times C}$ is first divided into a grid of N patches of size $P \times P$ where $= \frac{HW}{P^2}$. Each patch is flattened and projected into a D-dimensional embedding space through a linear layer as given Equation 5.

$$z_0^i = E \cdot flatten(x^i) + p_i, \qquad i = 1,\ 2, \ldots ., N \quad (5)$$

where, x^i is the i^{th} image patch, $E \in R^{D \times (P^2 \cdot C)}$ is the learnable patch embedding matrix, $P_i \in R^D$ is the learnable positional embedding added to each patch token to retain spatial information.

In addition, a learnable classification token $z_0^{[cls]} \in R^D$ is prepended to the patch sequence, which serves as the aggregated representation of the input image after processing through the Transformer layers (Kamal et al., 2025). The final input to the Transformer encoder is shown Equation 6.

$$Z_0 = \left[z_0^{[cls]},\ z_0^1,\ z_0^2, \ldots \ldots .,\ z_0^{N]} \right] \in R^{(N+1) \times D} \quad (6)$$

The Transformer encoder consists of L identical layers, each composed of a multi-head self-attention (MSA) mechanism followed by a position-wise feed-forward network (FFN). Each layer also includes residual connections and layer normalization as seen Equations 7 and 8 (Lu et al., 2025).

$$\widehat{Z}_l = MSA(LN(Z_{l-1})) + Z_{l-1} \quad (7)$$

$$Z_l = FNN(LN(\widehat{Z}_l)) + \widehat{Z}_l, \qquad l = 1, \ldots ., L \quad (8)$$

Here, the MSA operation splits the input into h heads and performs scaled dot-product attention in parallel as given in Equation 9.

$$Attention(Q,\ K,\ V) = \ softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (9)$$

where $Q,\ K,\ V \in R^{(N+1) \times d_k}$ are the query, key, and value matrices computed from the input via learned linear projections, and $d_k = {}^{D}/_{h}$ is the dimensionality of each head. The *softmax* function transforms the similarity scores into a probability distribution as given the Equation 10.

$$softmax(a_i) = \frac{e^{a_i}}{\sum_{j=1}^{n} e^{a_i}} \qquad i = 1, 2, \ldots, n \quad (10)$$

The ViT backbone used in this study was based on the "ViT-Base" variant, configured with the following parameters: patch size P=16, embedding dimension D=768, number of transformer layers L=12, number of attention heads h=12, feed-forward dimension d_{ff}=3072.

To further enrich the discriminative capability of ViT features, we introduced a FAM at the output of the transformer. This module, similar to the one used in ConvNeXt, emphasizes important channels in the output embedding of the classification token $z_L^{[cls]}$, based on global channel context. Given the final Transformer output Z_L, the class token vector $z_L^{[cls]} \in R^D$ is passed through a SE-inspired gating mechanism as shown in Equations 11 and 12 (Padshetty and Umashetty, 2024).

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z_L^{[cls]})) \quad (11)$$

$$\hat{z}_L^{[cls]} = s \cdot z_L^{[cls]} \quad (12)$$

This attention-weighted representation $z_L^{[cls]}$ captures the globally aggregated and recalibrated semantic information, which is later fused with the multiscale ConvNeXt features during the dynamic fusion stage of CMDF-Net as seen Figure 4. Also, the global modeling capacity of ViT given as Figure 6 robust feature extraction across both local textures and global structures in diseased mango leaf images.

3.5 Dynamic feature fusion module

The Dynamic Feature Fusion Module (DFFM) was specifically designed to effectively integrate the complementary strengths of ConvNeXt and ViT backbones within the proposed CMDF-Net architecture. While ConvNeXt provides rich local representations through hierarchical convolutional processing, ViT contributes global contextual dependencies via self-attention mechanisms (Duan et al., 2025). However, naive concatenation or addition of features from these heterogeneous sources may result in sub-

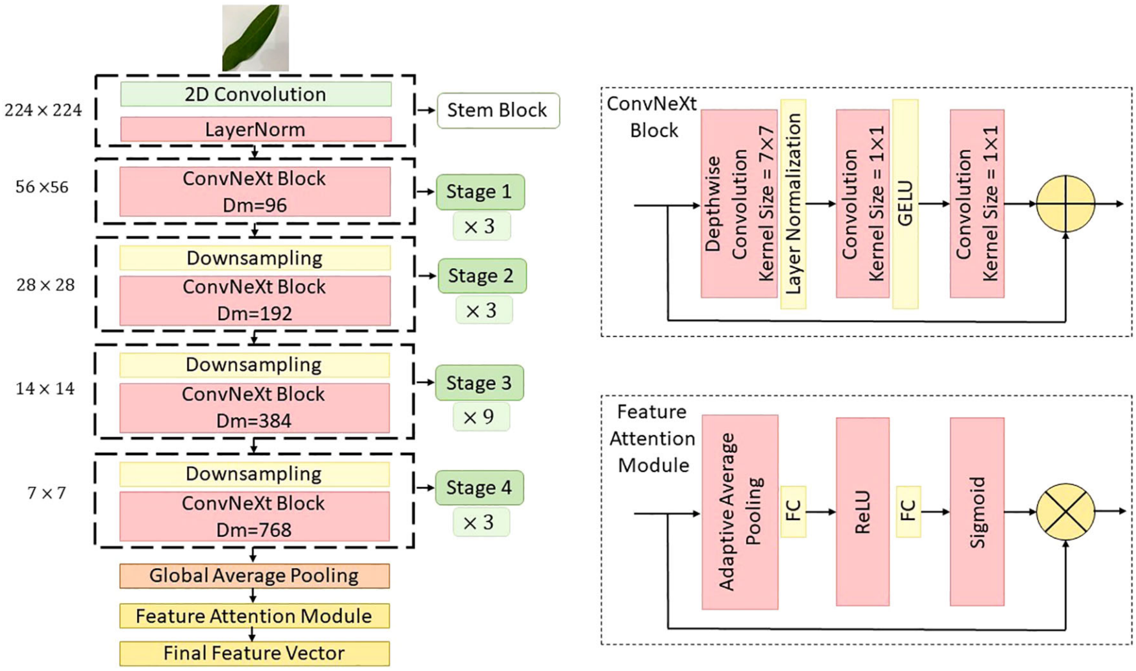


FIGURE 5
Detailed schematic of the ConvNeXt backbone as implemented within the MangoLeafCMDF-FAMNet framework, showing key stages of convolutional feature extraction and attention recalibration.

optimal representations due to mismatched semantics and scale. Therefore, DFFM aims to learn adaptive fusion weights that dynamically recalibrate and align the semantic contributions from both streams before final classification.

Let $F_{covn} \in R^{C\times H'\times W'}$ denote the multiscale feature map extracted from the ConvNeXt backbone after the final FAM, and $\hat{z}_{vit} \in R^{D}$ be the ViT-encoded class token vector refined by its respective FAM. To enable a joint fusion, the vector $\hat{z}_{vit}$ is first

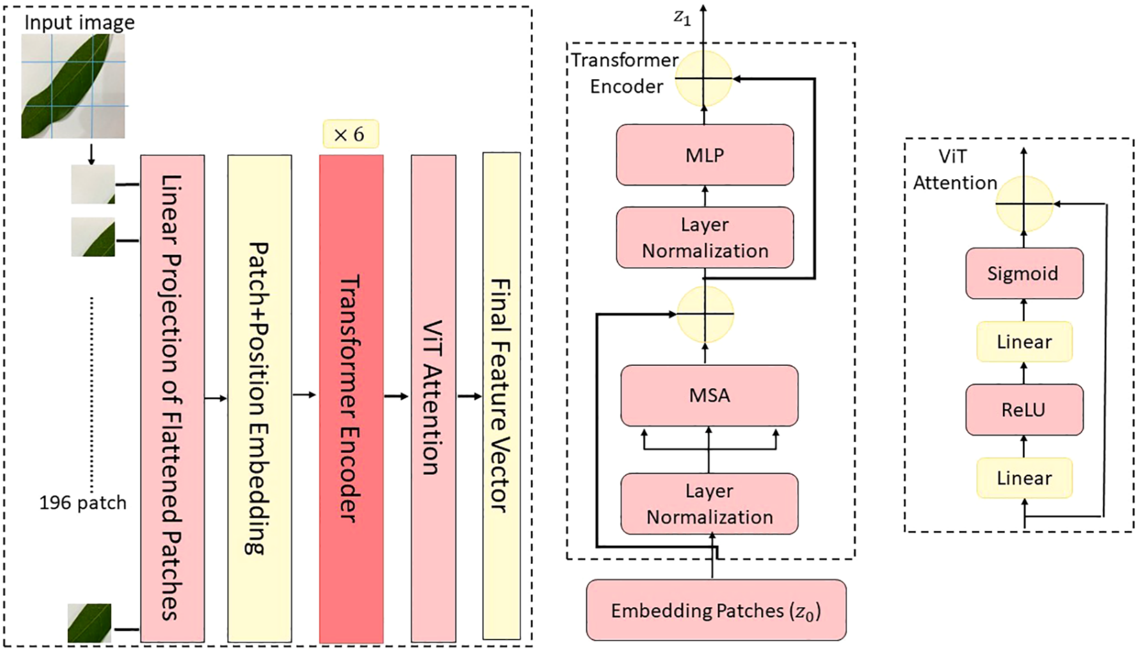


FIGURE 6
Architectural illustration of the ViT module employed in the proposed model, depicting patch embedding, transformer encoding, and output token generation stages.

spatially expanded and reshaped to match the spatial dimensions of F_{covn}, resulting in $F_{vit} \in R^{C\times H'\times W'}$ as shown in Equation 13 where a learnable linear transformation aligns D and C (Li et al., 2025).

$$F_{vit} = reshape(W_v \cdot \hat{z}_{vit} + b_v), \qquad W_v \varepsilon R^{C\times D} \tag{13}$$

Next, both feature maps are concatenated along the channel dimension to form a joint representation as Equation 14.

$$F_{joint} = concat(F_{covn}, \; F_{vit}) \in R^{2C\times H'\times W'} \tag{14}$$

A gated attention mechanism is used to model the interdependencies between the ConvNeXt and ViT features. The joint feature map is passed through a squeeze operation using global average pooling, followed by a two-layer fully connected network with non-linear activations as provided by Equation 15 (Duan et al., 2025). Where $W_1 \in R^{\frac{2C}{r}\times 2C}$, $W_2 \in R^{\frac{2C}{r}\times 2C}$ and $r = 16$.

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot GAP(F_{joint}))) \tag{15}$$

The resulting channel-wise attention vector $s \in R^{2C}$ acts as a set of dynamic fusion weights, controlling the contribution of each channel. This vector is split into two components corresponding to the original feature sources as provided by Equation 16. These weights are then used to recalibrate the characteristics of each branch as described in Equation 17.

$$s_{covn}, \; s_{vit} \in R^{C}, \qquad s = [s_{covn}; \; s_{vit}] \tag{16}$$

$$\hat{F}_{covn} = s_{covn} \odot F_{covn}, \qquad \hat{F}_{vit} = s_{vit} \odot F_{vit} \tag{17}$$

Finally, the recalibrated features are fused via element-wise summation to obtain the final feature representation as described in Equation 18.

$$F_{fused} = \hat{F}_{covn} + \hat{F}_{vit} \tag{18}$$

The fused feature map $F_{fused} \in R^{C\times H'\times W'}$ is passed through a global average pooling (GAP) layer, followed by a final fully connected classification head to predict the disease class label. This dynamic and learnable fusion strategy enables CMDF-Net to adaptively emphasize the most informative modalities depending on the content of each input image. The gating mechanism ensures that disease-specific patterns, whether localized or distributed globally, are optimally weighted, thereby enhancing the robustness and accuracy of the classification process.

3.6 Training strategy and evaluation protocol

The training strategy of the proposed MangoLeafCMDF-FAMNet model was meticulously designed to ensure stable convergence, optimal generalization, and fair performance evaluation across all experimental scenarios. All experiments were conducted using the PyTorch DL framework, ensuring efficient handling of high-dimensional image data and deep architectural components. Prior to training, all mango leaf images were resized to a spatial resolution of 224 × 224 pixels to ensure compatibility with the input dimensions of both ConvNeXt and ViT backbones. The ConvNeXt and ViT modules within CMDF-Net were initialized with pretrained weights from ImageNet-1K to leverage generic image feature representations. All additional layers—including FAM, DFFM, and the final classification head—were initialized using Kaiming He initialization for ReLU-based layers and Xavier initialization for linear projections, ensuring stable weight distribution at the start of training.

The model was trained using the AdamW optimizer, which combines adaptive gradient updates with decoupled weight decay regularization. The initial learning rate was set to 5×10^{-4} with a cosine annealing scheduler to facilitate smooth convergence. A warm-up phase of 10 epochs was employed, during which the learning rate was linearly increased from 1×10^{-5}. The weight decay was fixed at 1×10^{-4}, and a mini-batch size of 32 was used throughout the training. The cross-entropy loss function was utilized to compute the classification loss as described in Equation 19 (Zhou et al., 2019). Where y_i is the ground-truth label and $\hat{y}_i$ is the softmax probability of the predicted class for the i^{th} sample.

$$L_{CE} = -\sum_{i=1}^{N} y_i log(\hat{y}_i) \tag{19}$$

Each model was trained for a maximum of 100 epochs. However, early stopping with a patience value of 15 epochs was employed based on the validation loss to prevent overfitting and unnecessary computations. To ensure reliable and unbiased performance evaluation, 5-FCVP was performed. In each fold, the dataset was split into training (80%) and validation (20%) sets, maintaining class distribution. The average of all five folds was reported for each evaluation metric. To comprehensively evaluate the effectiveness of the proposed method, multiple performance metrics were employed: CA, RCL, PRC, MCC, and κ. These metrics collectively provide insights into the model's overall predictive power, class-wise sensitivity, balance, and inter-rater agreement, respectively. CA indicates the proportion of correctly classified samples out of the total number of instances. It is calculated as described in Equation 20 (Yavuz and Aydemir, 2016).

$$CA = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

where TP, TN, FP, and FN denote the true positives, true negatives, false positives, and false negatives, respectively. RCL, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify positive instances as explained in Equation 21. PRC reflects the proportion of true positive predictions among all positive predictions made by the model as shown in Equation 22 (Ergün, 2024).

$$RCL = \frac{TP}{TP + TN} \tag{21}$$

$$PRC = \frac{TP}{TP + TPP} \tag{22}$$

TABLE 4 Classification results of MangoLeafCMDF-FAMNet on the MLD1, MLD2, and MLD3 across 5- FCVP.

Datasets	Fold	Metrics				
		CA	RCL	PRC	MCC	Kappa
MLD1	Fold 1	0.9992	0.9992	0.9992	0.9991	0.9991
	Fold 2	0.9977	0.9976	0.9978	0.9973	0.9973
	Fold 3	1.0000	1.0000	1.0000	1.0000	1.0000
	Fold 4	0.9938	0.9937	0.9937	0.9929	0.9929
	Fold 5	0.9984	0.9985	0.9985	0.9982	0.9982
MLD2	Fold 1	0.9980	0.9979	0.9980	0.9975	0.9975
	Fold 2	0.9970	0.9969	0.9972	0.9963	0.9962
	Fold 3	1.0000	1.0000	1.0000	1.0000	1.0000
	Fold 4	0.9990	0.9991	0.9989	0.9988	0.9987
	Fold 5	1.0000	1.0000	1.0000	1.0000	1.0000
MLD3	Fold 1	0.9949	0.9956	0.9960	0.9941	0.9941
	Fold 2	0.9918	0.9932	0.9923	0.9905	0.9904
	Fold 3	0.9937	0.9945	0.9941	0.9927	0.9927
	Fold 4	0.9945	0.9953	0.9950	0.9936	0.9936
	Fold 5	0.9965	0.9970	0.9967	0.9959	0.9959

MCC is a robust measure that takes into account all four elements of the confusion matrix and is especially valuable for imbalanced datasets as described in Equation 23 (Rozenfeld et al., 2024). It returns a value between -1 and 1, where 1 indicates perfect prediction, 0 means no better than random guessing, and -1 represents total disagreement.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{23}$$

Kappa evaluates the agreement between predicted and actual classifications, adjusted for chance. It is defined as given in the Equation 24 (Ergün and Aydemir, 2020). Where p_0 is the observed agreement and p_e is the expected agreement by random chance.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{24}$$

3.7 Classification head

The classification head module serves as the terminal decision-making component of the CMDF-Net architecture, synthesizing the high-level, semantically rich features obtained from the dynamically fused ConvNeXt and ViT representations. Its primary objective is to project the fused feature map into a low-dimensional space corresponding to the number of disease categories and produce the final class probabilities through a softmax activation function. We denote the output feature tensor generated by the DFF module as $F_{fused} \in R^{R \times W \times C}$, where H, W, and C represent the spatial height, width, and number of channels of the fused feature map, respectively. Before classification, global spatial information is condensed using a GAP operation as described in Equation 25 (Hsiao et al., 2019).

$$z = GAP(F_{fused}) \in R^{C} \tag{25}$$

This operation ensures translational invariance and reduces the number of trainable parameters by eliminating the need for fully connected layers at the spatial level. The pooled vector z is then passed through a fully connected (FC) layer followed by a softmax function to obtain the final class probabilities as explained in Equation 26.

$$\hat{y} = softmax(W_c z + b_c) \tag{26}$$

where $W_c \in R^{K \times C}$ and $b_c \in R^{K}$ denote the weight matrix and bias vector of the classification layer, and K is the number of target classes. To enhance the expressiveness of the classification head while maintaining generalization, dropout regularization with a rate of 0.3 was employed prior to the final linear layer. This stochastic regularization strategy helps mitigate overfitting by randomly deactivating neurons during training.

4 Results

In this study, the proposed MangoLeafCMDF-FAMNet architecture was rigorously evaluated using a 5-FCVP across three publicly available datasets: MLD1, MLD2, and MLD3. The training phase employed the AdamW optimizer with an initial learning rate set to 0.00005 and a weight decay of 0.0001. Cross-entropy loss was used as the objective function to guide the optimization process. Model performance was assessed comprehensively using multiple evaluation metrics, namely CA, RCL, PRC, MCC, and kappa. In addition, confusion matrices and high-dimensional feature distributions visualized through t-SNE were generated to provide further insights into the discriminative capability of the model. Table 4 summarizes the performance metrics obtained for each fold across all three datasets. For MLD1, the model achieved remarkably high performance, consistently exceeding 99.00% CA across all folds. Specifically, Fold 3 yielded a perfect CA, RCL, and PRC of 1.0000, with corresponding MCC and kappa of 1.0000, indicating flawless classification without any mispredictions. Even in the comparatively lower-performing Fold 4, MangoLeafCMDF-FAMNet still maintained an outstanding CA of 0.9938, demonstrating its robustness against potential variability in the data splits. Similarly, for MLD2, the model maintained exceptional performance. Perfect scores were achieved in Folds 3 and 5, mirroring the trends observed in MLD1. Notably, the lowest CA across all folds was 0.9970, which still reflects a near-perfect classification capability. The consistently high MCC and kappa across folds further underline the model's strong agreement between the predicted and true class labels, confirming its reliability. On MLD3, which is inherently more challenging due to greater symptom variability and inter-class similarity, MangoLeafCMDF-FAMNet continued to demonstrate excellent performance. The CA values across the five folds ranged from 0.9918 to 0.9965, with the highest score achieved in Fold 5. RCL and PRC closely mirrored the trends of CA, and the high MCC and kappa reaffirmed the model's ability to generalize well even under more complex conditions.

Following, a detailed class-wise performance analysis was conducted to further assess the robustness and generalization ability of MangoLeafCMDF-FAMNet. Specifically, RCL and PRC were calculated for each class across all folds on the MLD1, MLD2, and MLD3 datasets. For the MLD1 dataset, the model demonstrated outstanding classification capabilities. As shown in Figure 7, the average RCL values were 0.9988 for Anthracnose, 0.9988 for Bacterial Canker, 1.0000 for Cutting Weevil, 0.9987 for Die Back, 0.9951 for Gall Midge, 0.9987 for Healthy leaves, 0.9961 for Powdery Mildew, and 0.9962 for Sooty Mould. In terms of PRC, the averages were equally high, reaching 1.0000 for several classes, with minor reductions to 0.9962 and 0.9935 for Powdery Mildew and Sooty Mould, respectively. These results confirmed that the model could accurately distinguish subtle disease symptoms even under slight class imbalance or symptom similarity.

Similarly, in the MLD2 dataset, the model achieved nearly perfect classification. The average RCL scores were recorded at 0.9991 for Anthracnose, 1.0000 for Die Back, 0.9990 for Gall Midge, 1.0000 for Healthy leaves, and 0.9958 for Powdery Mildew. Correspondingly, the PRC values were consistently excellent, with an average exceeding 0.9990 for all categories. Notably, the model maintained strong performance even in folds where small fluctuations in Powdery Mildew recognition were observed, demonstrating resilience against minor dataset variations.

The analysis of the MLD3 dataset, which is inherently more challenging due to the larger number of disease classes, also revealed robust model performance. The average RCL values across folds were 0.9869 for Anthracnose, 0.9949 for Bacterial Canker, 0.9910 for Cutting Weevil, 0.9992 for Die Back, 0.9909 for Gall Midge, and 1.0000 for Healthy leaves, Powdery Mildew, and Sooty Mould. PRC values aligned closely, maintaining averages above 0.9850 for all categories. While slight drops were noted in classes such as Anthracnose and Sooty Mould, the model overall preserved an exceptional balance between sensitivity and specificity. To visually capture these findings, RCL and PRC radar plots were generated across all folds, as depicted in Figure 7. In these visualizations, Fold 1 is represented in dark blue, Fold 2 in orange, Fold 3 in gray, Fold 4 in yellow, Fold 5 in dark navy, and the average of all folds is plotted in green.

Furthermore, to comprehensively evaluate the class-level robustness of MangoLeafCMDF-FAMNet, class-wise MCC and Kappa were calculated and illustrated in Figure 8 for the MLD1 dataset across the 5-FCVP. Specifically, MCC scores remained extremely high for all disease categories, often reaching perfect agreement across most folds. Minor variations were observed only in the Gall Midge and Sooty Mould classes, where the MCC values slightly dropped but still remained above 0.98, demonstrating the strong generalization capacity of the model without signs of overfitting. Similarly, the Kappa mirrored the MCC trends, with near-perfect agreement across all folds and classes. On average, both MCC and Kappa exceeded 0.99 for nearly every class, highlighting the model's consistent ability to correctly classify diverse disease symptoms under varying validation conditions.

In addition, we further computed the MCC and Kappa for the MLD2, as visualized in Figure 9. As illustrated, the MCC scores for all classes consistently achieved near-perfect values, with almost every fold yielding scores of 1.0000 across the Anthracnose, Die Back, Healthy, and Powdery Mildew categories. A slight deviation was observed in the Gall Midge class during Fold 1 and Fold 2, where the MCC values dropped marginally but still remained exceedingly high, thereby underscoring the model's remarkable stability even in the presence of subtle intra-class variations. Similarly, the Kappa mirrored these trends, maintaining values close to 1.0000 across all classes and folds, reaffirming the excellent agreement between predicted and true labels. The minimal variability observed in the Gall Midge class reflects realistic complexities inherent in agricultural imaging datasets, yet the exceedingly high average scores across all classes strongly indicate that MangoLeafCMDF-FAMNet successfully mitigates overfitting and maintains robust generalization.

Furthermore, to ensure a comprehensive performance analysis, we computed the class-wise MCC and Kappa of the MangoLeafCMDF-FAMNet on the MLD3, as presented in Figure 10. The results demonstrated that the proposed model consistently achieved exceptionally high MCC values across all classes and folds. Specifically, MCC scores for classes such as Die Back, Healthy, Powdery Mildew, and Sooty Mould remained at or extremely close to 1.0000 across all folds. Although slight variations were noted in classes like Anthracnose and Bacterial Canker, the MCC scores still hovered around 0.98–0.99, reflecting highly reliable performance even in more challenging classes. Similarly, the Kappa closely followed the MCC trends, indicating outstanding agreement between predicted and ground truth labels across all folds and classes.

Furthermore, the feature distributions fused by MangoLeafCMDF-FAMNet were qualitatively analyzed using t-SNE, as illustrated in Figure 11 for each dataset individually. t-SNE serves as a powerful non-linear dimensionality reduction technique that projects high-dimensional feature representations into a two-dimensional space, facilitating the visualization of complex feature relationships. The t-SNE plots reveal that features extracted by MangoLeafCMDF-FAMNet exhibit clear, compact, and well-separated clusters for different disease classes across all three datasets. This visual evidence supports the numerical findings, indicating that the model effectively learns discriminative and disease-specific representations without significant overlap among categories. Moreover, the distinct cluster formations further demonstrate the absence of overfitting, affirming the model's strong generalization capability to unseen samples.

The average confusion matrices for the MLD1, MLD2, and MLD3 datasets, computed across the 5-FCVP and illustrated in Figure 12, provide critical insights into the class-specific discriminative capabilities of MangoLeafCMDF-FAMNet. These matrices (vertical axis: true labels; horizontal axis: predicted

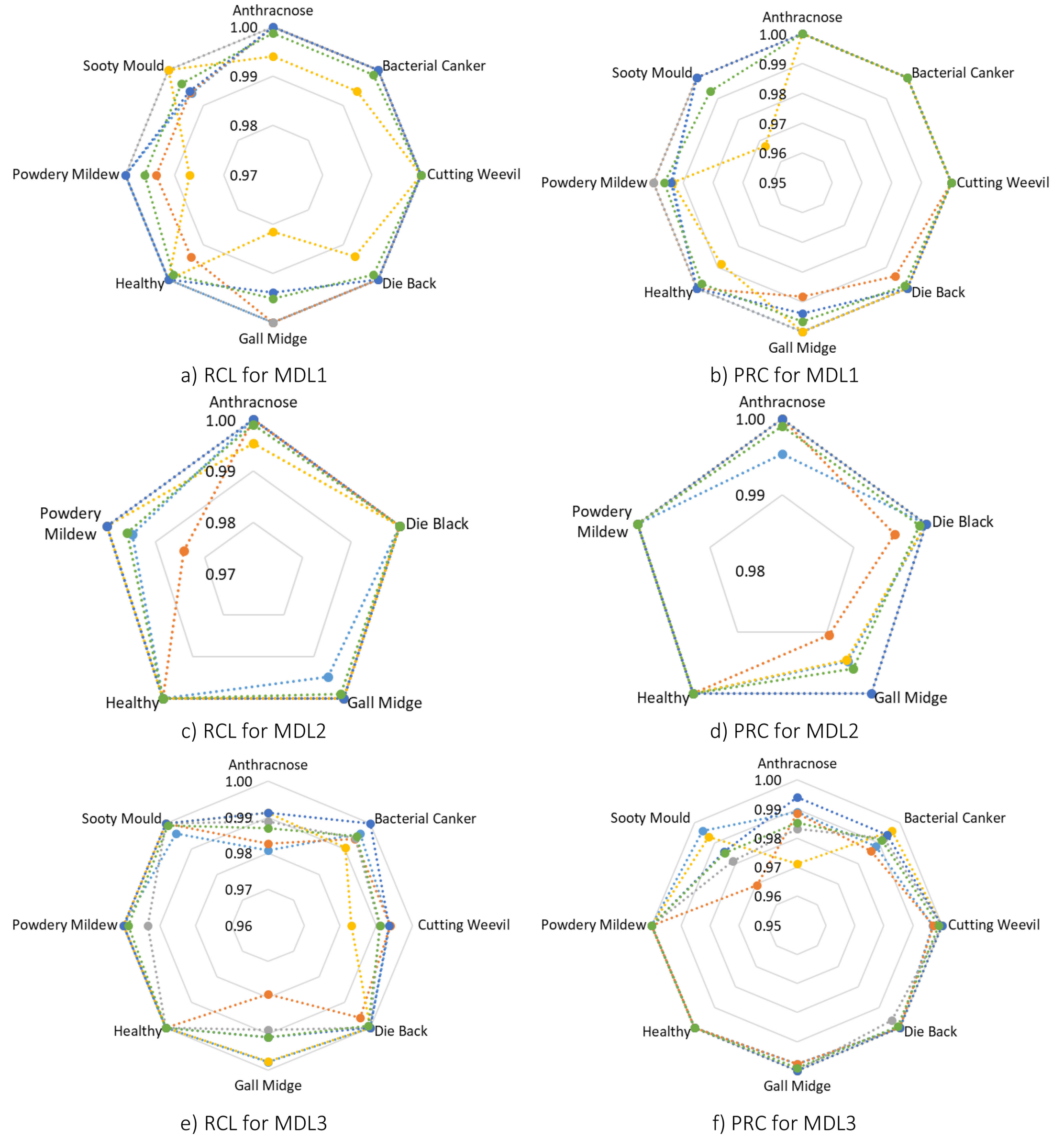


FIGURE 7
Radar plots visualizing class-wise RCL and PRC metrics for MangoLeafCMDF-FAMNet across all five folds of the 5-FCVP on MLD1, MLD2, and MLD3 datasets. Each fold is color-coded, with the green line representing the average across folds **(a)** RCL for MLD1, **(b)** PRC for MLD1, **(c)** RCL for MLD2, **(d)** PRC for MLD2, **(e)** RCL for MLD3, and **(f)** PRC for MLD3.

labels) reveal near-perfect diagonal dominance, underscoring the model's ability to minimize misclassifications while maintaining high intra-class consistency. For MLD1, the matrix demonstrated exceptional precision, with Anthracnose and Bacterial Canker—classes often confused due to overlapping lesion patterns—achieving 159.80 correct predictions, respectively. Only minor off-diagonal errors were observed: 0.20 of Anthracnose samples were misclassified as Sooty Mould, while 0.20 of Bacterial Canker cases were incorrectly assigned to Healthy. The Cutting Weevil class exhibited flawless performance, with all 160 samples correctly identified. Similarly, Die Back and Gall Midge achieved near-perfect classification, with diagonal values of 159.80 and 159.40, respectively.

In MLD2, the matrix highlighted robust performance under increased symptom variability. Gall Midge, a class with subtle morphological features, achieved 199.80 correct predictions, with only 0.20 confusion with Healthy leaves. Die Back and Healthy were resolved with high precision, as evidenced by diagonal entries of 200.00. The most notable misclassification occurred in the Powdery Mildew class, where 0.60 samples were erroneously predicted as Gall Midge, likely due to shared textural patterns in late-stage infections.

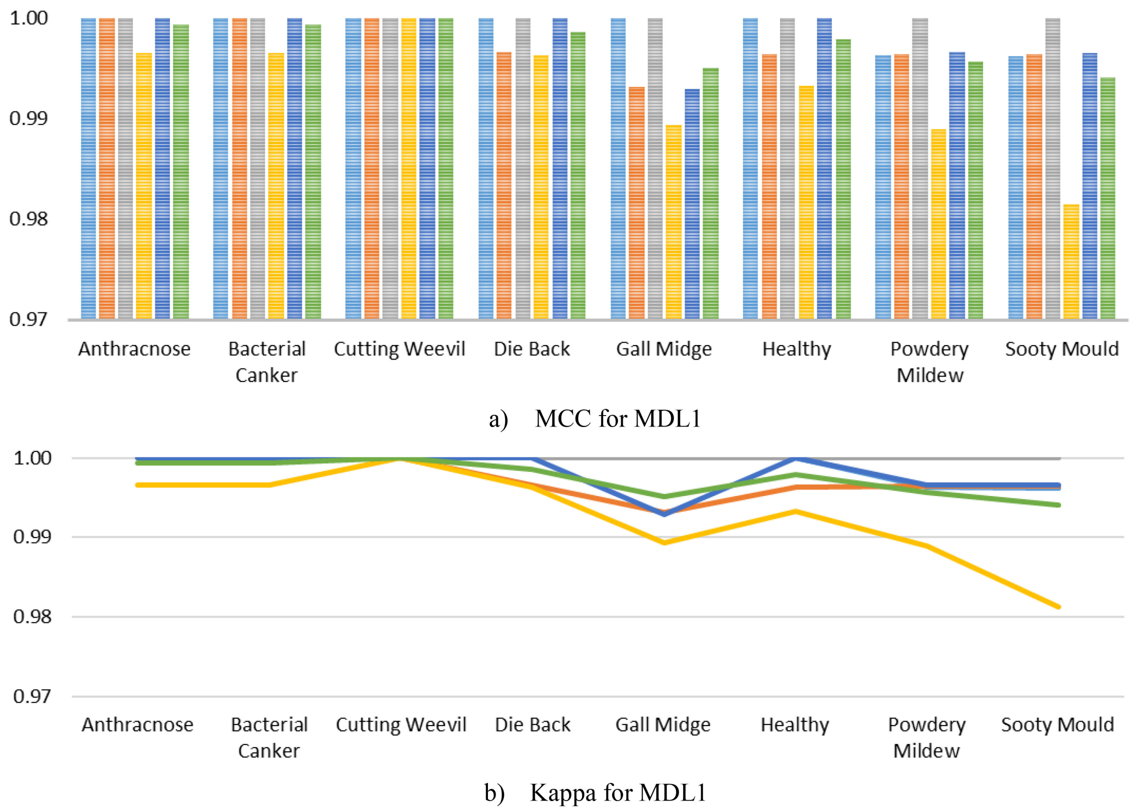


FIGURE 8
Per-class MCC and Kappa metrics achieved by MangoLeafCMDF-FAMNet for the MLD1 dataset, evaluated over the 5-FCVP. Colored lines indicate individual folds, while the green line shows the average performance **(a)** MCC for MLD1 and **(b)** Kappa for MLD1.

For MLD3, the matrix further validated the model's generalization capability. The Die Back class, often prone to misclassification due to its overlapping early-stage symptoms with Gall Midge, achieved a strong diagonal value of 255.80 correct predictions, reflecting the model's ability to discern subtle differences in lesion distribution. Powdery Mildew, a class with visually ambiguous fungal patterns, was correctly classified in 154.80 instances, with only 0.20 samples misassigned to Die Back—a negligible error likely attributable to shared textural features in advanced infection stages. Gall Midge, despite its complex morphological variations across growth cycles, demonstrated exceptional performance with 442.60 accurate predictions. A minimal leakage of 3.80 samples to Sooty Mould was observed, potentially stemming from similarities in necrotic patterning under low-light imaging conditions. The Healthy class once again exhibited flawless discriminative capability, with all 250.00 samples correctly identified, underscoring the model's precision in isolating disease-specific features from healthy tissue. Sooty Mould, a class frequently confused with Powdery Mildew in conventional methods, achieved a near-perfect diagonal score of 264.80, further highlighting the architecture's proficiency in resolving spectral ambiguities. These results collectively affirm that MangoLeafCMDF-FAMNet generalizes robustly across diverse data distributions, making it particularly suitable for real-world agricultural applications where symptom variability and class overlap are prevalent.

In order to robustly validate the effectiveness of the proposed MangoLeafCMDF-FAMNet, an extensive comparative analysis was conducted against prominent baseline models, including MangoLeafCMDF-Net, ViT, and ConvNeXt, across the MLD1, MLD2, and MLD3 datasets. The results, summarized in Table 5, clearly demonstrate the superiority of MangoLeafCMDF-FAMNet across all evaluation metrics. Specifically, on MLD1, MangoLeafCMDF-FAMNet achieved a CA of 0.9978, outperforming MangoLeafCMDF-Net with 0.9961, ConvNeXt with 0.9939, and ViT with a considerably lower score of 0.9256. In terms of RCL and PRC, MangoLeafCMDF-FAMNet consistently achieved 0.9978 for both, substantially higher than the competing models. On MLD2, the proposed model maintained its leading performance, reaching a CA of 0.9988, while MangoLeafCMDF-Net, ConvNeXt, and ViT achieved 0.9960, 0.9814, and 0.9176, respectively. Even on the more challenging MLD3 dataset, MangoLeafCMDF-FAMNet maintained its superiority, recording a CA of 0.9943, whereas ConvNeXt achieved 0.9864 and ViT remained at 0.9111. When considering robustness metrics, MangoLeafCMDF-FAMNet attained an MCC of 0.9975 and a

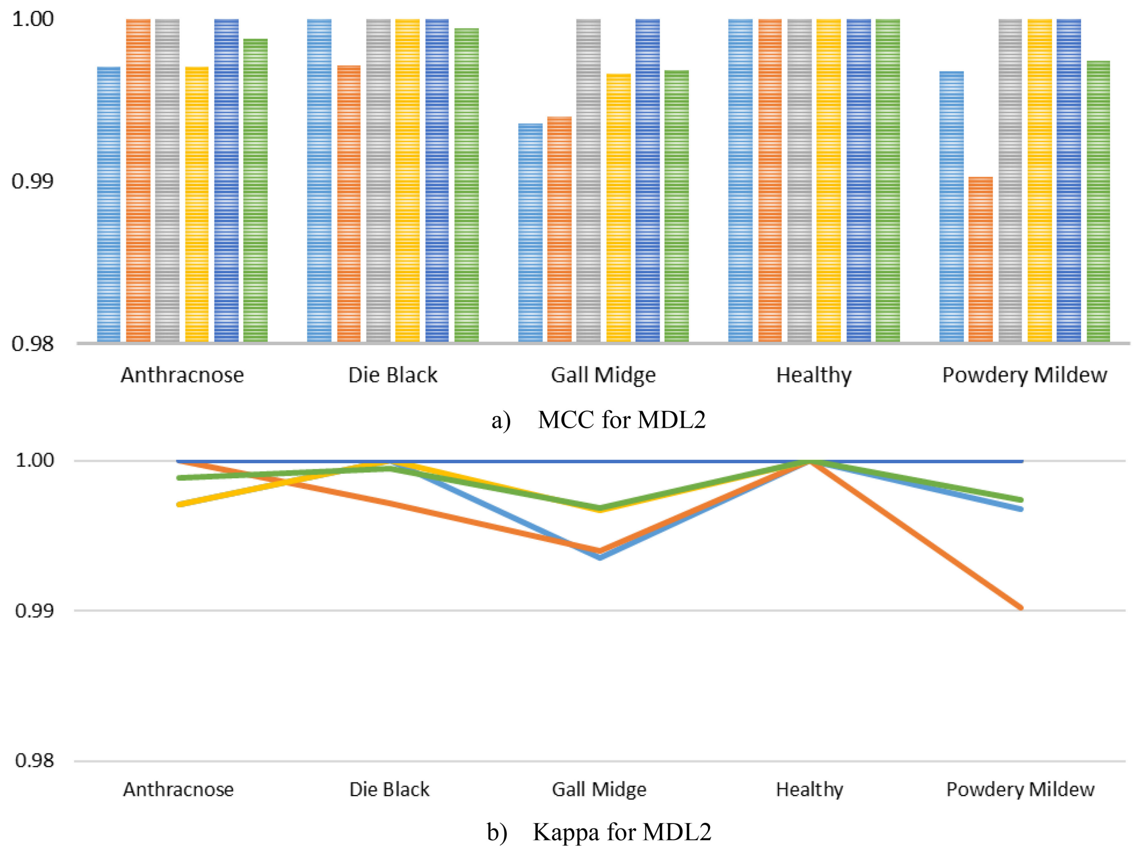


FIGURE 9
Class-wise analysis of **(a)** MCC and **(b)** Kappa obtained from MangoLeafCMDF-FAMNet on the MLD2 dataset under 5-FCVP. Individual folds are color-coded; the green line indicates the averaged result.

Kappa of 0.9975 on MLD1, again consistently exceeding those of the other methods. This outstanding performance was mirrored across MLD2 and MLD3, highlighting not only accuracy but also high reliability and class-wise agreement.

To rigorously evaluate whether the performance differences between the proposed MangoLeafCMDF-FAMNet and baseline architectures were statistically significant, we conducted a Tukey Honest Significant Difference (HSD) *post-hoc* test based on the CA values obtained across five folds for each dataset. The results revealed that MangoLeafCMDF-FAMNet significantly outperformed the ViT model across all three datasets (MLD1, MLD2, and MLD3), with adjusted *p*-values consistently below 0.01. Likewise, when compared to the ConvNeXt backbone, the proposed method demonstrated statistically significant improvements in MLD1 ($p = 0.024$) and MLD3 ($p = 0.038$), while achieving a strong yet marginally nonsignificant advantage in MLD2 ($p = 0.067$). These findings underscore the consistent superiority of MangoLeafCMDF-FAMNet in terms of CA, particularly highlighting the added value of its attention-enhanced, cross-modal fusion strategy. Furthermore, the absence of overlapping confidence intervals supports the robustness of the proposed model's improvements and suggests that the observed performance gains are unlikely due to random variation or overfitting.

5 Conclusion

This study introduced MangoLeafCMDF-FAMNet, a novel attention-augmented hybrid DL architecture tailored for robust multi-class mango leaf disease classification. By synergistically combining ConvNeXt and ViT backbones through a CMDF strategy, and further enhancing their output with FAMs, the proposed method effectively captured both fine-grained local patterns and global semantic context. Extensive experiments on three publicly available datasets (MLD1, MLD2, and MLD3) demonstrated the model's superior classification performance across various evaluation metrics, including CA, RCL, PRC, MCC, and Kappa.

The proposed architecture consistently achieved exceptionally high accuracies across all datasets, reaching 0.9978 on MLD1, 0.9988 on MLD2, and 0.9943 on MLD3. These results notably outperformed competing baselines such as ViT, ConvNeXt, and MangoLeafCMDF-Net. Class-wise evaluations further confirmed

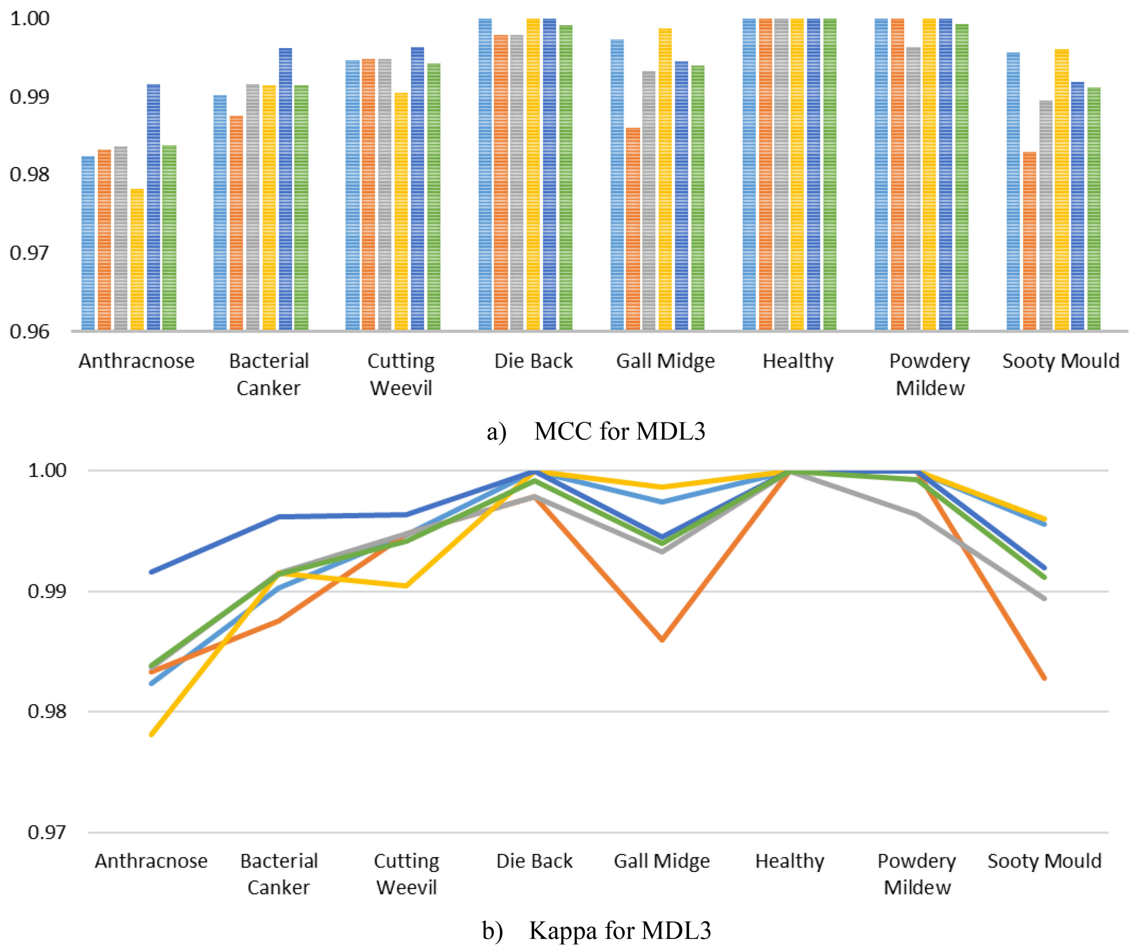


FIGURE 10
Visualization of per-class **(a)** MCC and **(b)** Kappa for MangoLeafCMDF-FAMNet on the MLD3 dataset, based on 5-FCVP. Fold-wise trends and average values are presented to demonstrate performance consistency.

the model's capacity to distinguish complex and visually similar disease symptoms with high RCL and PRC. Visualizations via t-SNE and confusion matrices affirmed the learned feature separability and robustness against inter-class confusion, while the statistical analyses via Tukey HSD *post-hoc* testing verified that the observed improvements were statistically significant ($p < 0.05$) in most comparisons, particularly over the ViT and ConvNeXt baselines.

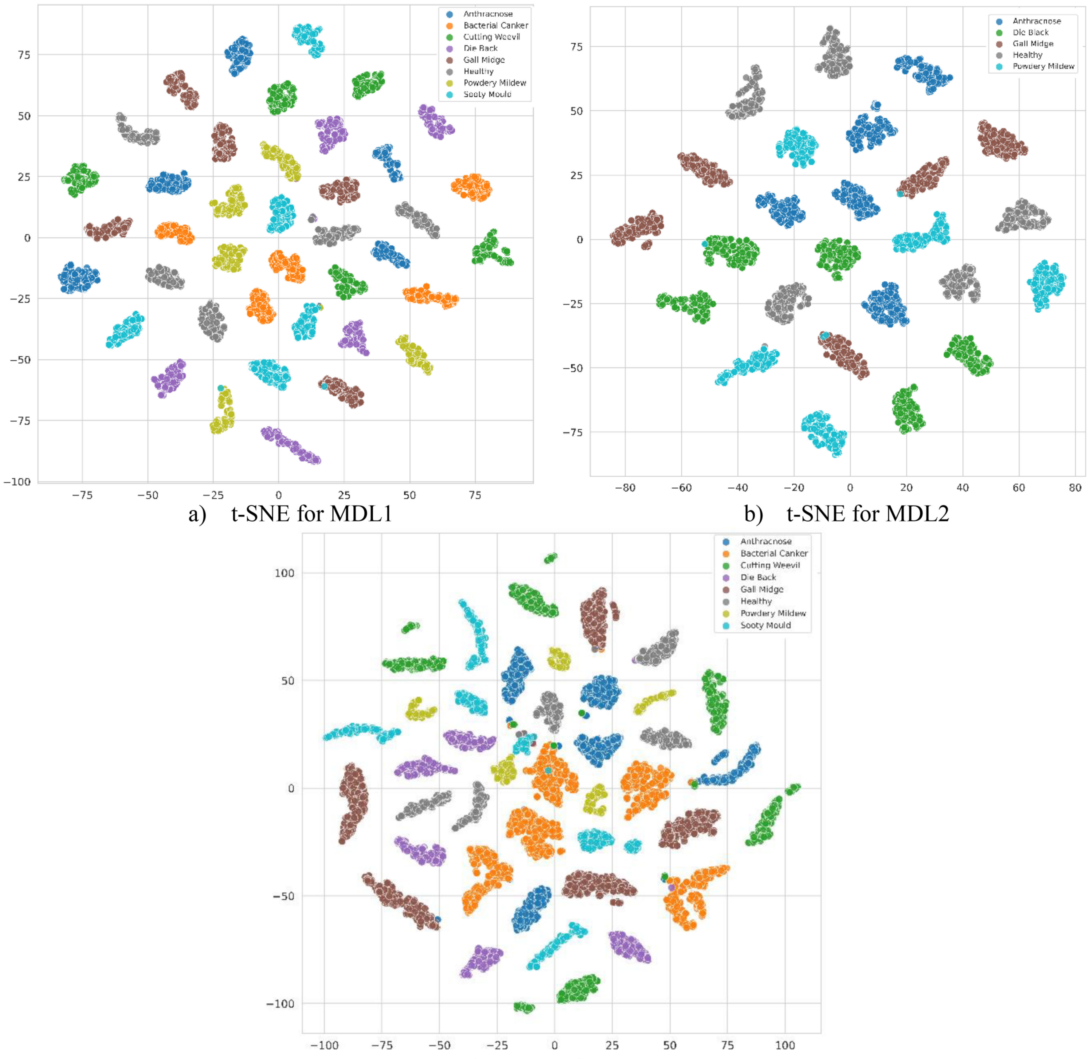
Furthermore, the architecture demonstrated a remarkable ability to generalize without overfitting, even on MLD3—a dataset characterized by greater class imbalance and symptom variability. The inclusion of FAMs was instrumental in adaptively amplifying disease-relevant features while suppressing irrelevant or redundant information, thereby enhancing class separability across diverse visual domains.

Despite these promising outcomes, the study is not without limitations. First, although the model was tested across three comprehensive datasets, all samples were derived from controlled imaging conditions. Future work should explore the model's applicability to in-field images collected under varying lighting, occlusion, and background clutter. Second, while the proposed model achieved excellent results in disease identification, it currently does not support disease severity estimation, which is crucial for more nuanced decision-making in real-world scenarios.

As future research directions, we aim to extend the MangoLeafCMDF-FAMNet architecture for real-time mobile deployment in smart agriculture systems, incorporate multimodal inputs such as hyperspectral or thermal imagery to improve resilience under environmental variations, and explore the integration of explainability modules to foster model transparency for end-users such as farmers and agronomists. In conclusion, MangoLeafCMDF-FAMNet represents a scientifically grounded, practically scalable, and statistically validated advancement in automated plant disease recognition.

6 Discussion

The experimental findings obtained in this study demonstrate that MangoLeafCMDF-FAMNet offers a highly effective solution for the complex task of multi-class mango leaf disease classification. By integrating ConvNeXt and ViT within a unified hybrid



c) t-SNE for MDL3

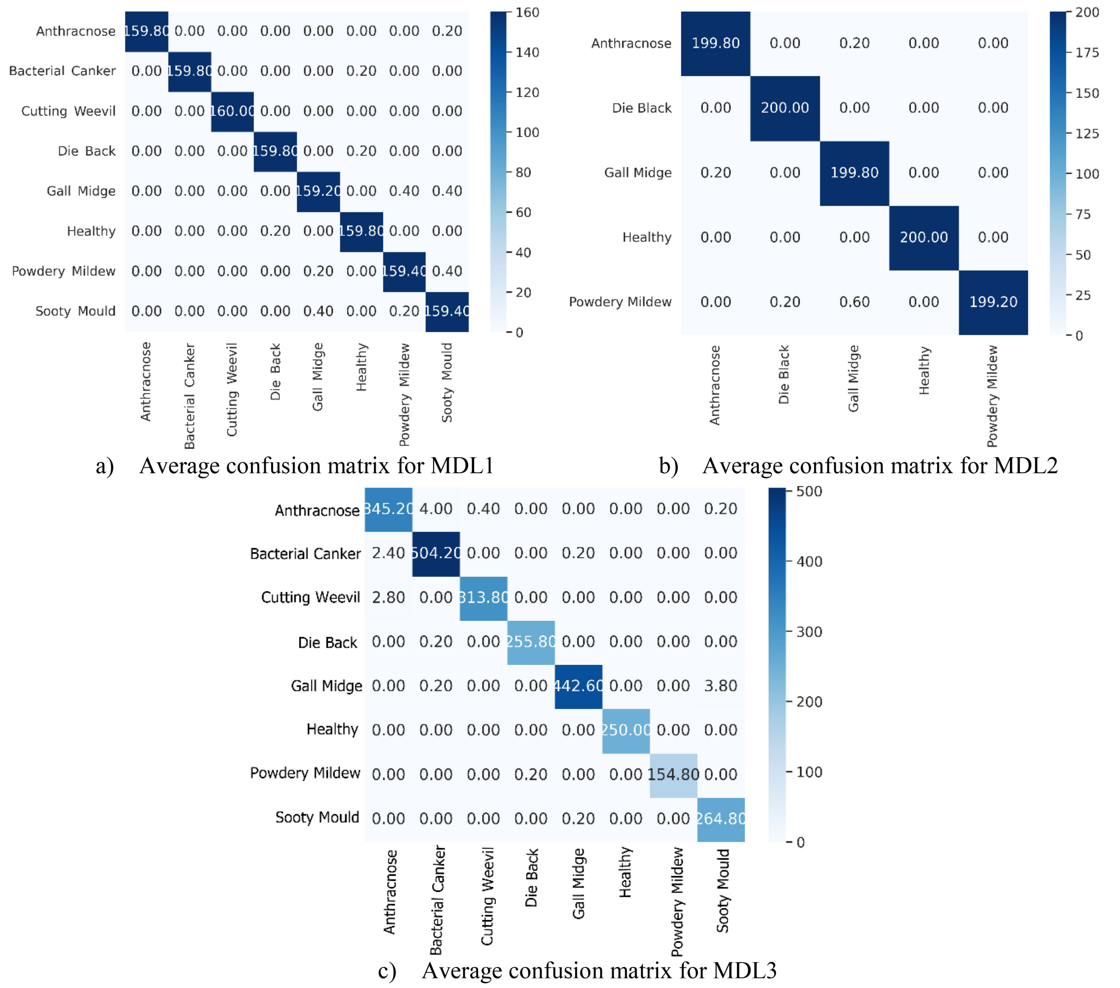
FIGURE 11
Two-dimensional t-SNE plots of the fused feature embeddings produced by MangoLeafCMDF-FAMNet across **(a)** MLD1, **(b)** MLD2, and **(c)** MLD3 datasets. The visualization illustrates the discriminative capacity and separability of learned features among disease classes.

architecture and augmenting their capabilities with FAM, the proposed model achieves a refined balance between local feature extraction and global semantic understanding. This synergy enables precise discrimination of disease types, particularly in cases where subtle morphological differences challenge conventional classifiers.

The model's high performance across three publicly available mango leaf datasets confirms its robustness and generalizability under controlled conditions. CA approaching 0.999, alongside strong MCC and kappa, indicate the architecture's capacity to produce stable, reliable, and interpretable predictions. Furthermore, the CMDF mechanism contributes significantly to the enrichment of feature representations, enabling more resilient learning from heterogeneous visual patterns.

Despite these strengths, it is essential to contextualize the results within the scope of the datasets utilized. All datasets in this study were collected under relatively uniform environmental settings, characterized by consistent lighting, minimal occlusions, and simplified backgrounds. While such conditions are favorable for model training and benchmarking, they may not fully reflect the variability encountered in operational agricultural environments. In practice, field images often contain challenges such as partial leaf visibility, shadowing, cluttered scenes, and inconsistent illumination, which may affect the model's generalization performance.

This observation highlights the importance of future work focused on validating the proposed framework using in-field image datasets collected in diverse and uncontrolled environments. Incorporating real-world variability into the training and evaluation pipeline will facilitate the development of more adaptive and field-deployable models. In addition, real-time image acquisition technologies, such as mobile devices and drone



a) Average confusion matrix for MDL1

b) Average confusion matrix for MDL2

c) Average confusion matrix for MDL3

FIGURE 12
Averaged confusion matrices of MangoLeafCMDF-FAMNet for the **(a)** MLD1, **(b)** MLD2, and **(c)** MLD3 datasets, computed over 5-FCVP. The vertical axis denotes the actual class labels, while the horizontal axis shows the predicted class labels.

platforms, present promising avenues for extending the system toward scalable agricultural decision support.

Another important consideration relates to the clinical utility of disease severity estimation. While the present model effectively identifies the disease type, it does not explicitly address the severity or progression stage of the infection. In real-world agricultural applications, the intensity of disease symptoms is a critical factor influencing treatment strategies and resource allocation. Thus, expanding the model to support ordinal or regression-based predictions for disease severity would significantly enhance its applicability. Although the datasets used in this study did not provide severity annotations, future efforts will focus on curating such datasets and developing models capable of jointly performing disease identification and severity grading.

Moreover, it is worth considering that the absence of diverse environmental conditions and severity-level annotations in the training data may limit the interpretability and practical utility of the current system. Addressing these challenges through targeted dataset development, attention to domain adaptation, and auxiliary prediction tasks will be essential for realizing the full potential of DL-based disease diagnosis in agricultural practice.

In conclusion, MangoLeafCMDF-FAMNet offers a robust and scalable architecture for automated mango leaf disease classification. By leveraging multi-level attention mechanisms and cross-modal fusion, the model provides a strong foundation for high-accuracy plant disease recognition. Future directions should emphasize improving real-world generalization and enhancing the interpretability of the system through disease severity assessment, ultimately supporting the broader goals of precision agriculture and sustainable crop management.

To evaluate the practical applicability of the proposed MangoLeafCMDF-FAMNet in real-world settings, we report key computational characteristics, including model complexity and inference efficiency. The model contains approximately 46.9 million trainable parameters, representing a balanced architectural design that ensures high discriminative power while maintaining computational feasibility. All training and evaluation experiments were performed using the PyTorch DL framework on a standard workstation equipped with an Intel(R) Core(TM) i7–9700 CPU and 8 GB of RAM, without

TABLE 5 Comparison of MangoLeafCMDF-FAMNet with MangoLeafCMDF-Net, ViT, and ConvNeXt across the MLD1, MLD2 and MLD3.

Metrics	MangoLeafCMDF-FAMNet			MangoLeafCMDF-Net			ViT			ConvNeXt		
	MLD1	MLD2	MLD3	MLD1	MLD2	MLD3	MLD1	MLD2	MLD3	MLD1	MLD2	MLD3
CA	0.9978	0.9988	0.9943	0.9961	0.9960	0.9919	0.9256	0.9176	0.9111	0.9939	0.9814	0.9864
RCL	0.9978	0.9988	0.9951	0.9961	0.9959	0.9933	0.9258	0.9173	0.9062	0.9939	0.9816	0.9861
PRC	0.9978	0.9988	0.9948	0.9962	0.9961	0.9929	0.9298	0.9219	0.9195	0.9940	0.9833	0.9884
MCC	0.9975	0.9985	0.9933	0.9956	0.9950	0.9906	0.9156	0.8984	0.8978	0.9931	0.9773	0.9843
Kappa	0.9975	0.9985	0.9933	0.9955	0.9950	0.9906	0.9150	0.8969	0.8967	0.9930	0.9767	0.9842

access to GPU acceleration. Under this configuration, the average inference time for a single 224×224-pixel image was observed to be approximately 180–200 milliseconds, depending on system load and batch scheduling. These results suggest that MangoLeafCMDF-FAMNet remains computationally viable even in resource-limited environments, which is particularly beneficial for field-deployable plant disease diagnosis systems where high-end hardware may not be available.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Author contributions

EE: Formal Analysis, Validation, Methodology, Supervision, Conceptualization, Writing – original draft, Software, Writing – review & editing, Visualization, Investigation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was financially supported by the Recep Tayyip Erdogan University Development Foundation (Grant number: 02025006018553). We sincerely appreciate their support, which contributed significantly to the completion of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alamri, F. S., Sadad, T., Almasoud, A. S., Aurangzeb, R. A., and Khan, A. (2025). Mango disease detection using fused vision transformer with convNeXt architecture. *Computers Materials Continua* 83, 1023–1039. doi: 10.32604/cmc.2025.061890

Arivazhagan, S., and Ligi, S. V. (2018). Mango leaf diseases identification using convolutional neural network. *Int. J. Pure Appl. Mathematics* 120, 11067–11079.

Bairwa, A. K., Singh, A., and Kumar, S. (2024). "Advances in mango leaf disease detection using deep neural networks," in *2024 international conference on modeling, simulation & Intelligent computing (MoSICom)* (Dubai, United Arab Emirates: IEEE), 75–80.

Chen, Y. C., Wang, J. C., Lee, M. H., Liu, A. C., and Jiang, J. A. (2024). Enhanced detection of mango leaf diseases in field environments using MSMP-CNN and transfer learning. *Comput. Electron. Agric.* 227, 109636. doi: 10.1016/j.compag.2024.109636

Duan, X., Liu, Y., You, Z., and Li, Z. (2025). Agricultural text classification method based on ERNIE 2.0 and multi-feature dynamic fusion. *IEEE Access.* 13, 52959–52971. doi: 10.1109/ACCESS.2025.3537277

Ergün, E. (2024). Deep learning-based multiclass classification for citrus anomaly detection in agriculture. *Signal Image Video Process.* 18, 8077–8088. doi: 10.1007/s11760-024-03025-6

Ergün, E. (2025). High precision banana variety identification using vision transformer based feature extraction and support vector machine. *Sci. Rep.* 15, 10366. doi: 10.1038/s41598-025-95466-0

Ergün, E., and Aydemir, Ö. (2020). A hybrid BCI using singular value decomposition values of the fast Walsh–Hadamard transform coefficients. *IEEE Trans. Cogn. Dev. Syst.* 15, 454–463. doi: 10.1109/TCDS.2020.3028785

Ford, J., Sadgrove, E., and Paul, D. (2025). Joint plant-spraypoint detector with ConvNeXt modules and HistMatch normalization. *Precis. Agric.* 26, 24. doi: 10.1007/s11119-024-10208-y

Fu, G., Lin, K., Lu, C., Wang, X., and Wang, T. (2025). Spindle thermal error regression prediction modeling based on ConvNeXt and weighted integration using thermal images. *Expert Syst. Appl.* 274, 127038. doi: 10.1016/j.eswa.2025.127038

Gautam, V., Ranjan, R. K., Dahiya, P., and Kumar, A. (2024). ESDNN: A novel ensembled stack deep neural network for mango leaf disease classification and detection. *Multimedia Tools Appl.* 83, 10989–11015. doi: 10.1007/s11042-023-16012-6

Hossain, M. A., Sakib, S., Abdullah, H. M., and Arman, S. E. (2024). Deep learning for mango leaf disease identification: A vision transformer perspective. *Heliyon* 10, e36361. doi: 10.1016/j.heliyon.2024.e36361

Hsiao, T. Y., Chang, Y. C., Chou, H. H., and Chiu, C. T. (2019). Filter-based deep-compression with global average pooling for convolutional networks. *J. Syst. Architecture* 95, 9–18. doi: 10.1016/j.sysarc.2019.02.008

Kamal, S., Sharma, P., Gupta, P. K., Siddiqui, M. K., Singh, A., and Dutt, A. (2025). DVTXAI: A novel deep vision transformer with an explainable AI-based framework and its application in agriculture. *J. Supercomputing* 81, 1–32. doi: 10.1007/s11227-024-06494-y

Li, S., Chen, Z., Xie, J., Zhang, H., and Guo, J. (2025). PD-YOLO: A novel weed detection method based on multi-scale feature fusion. *Front. Plant Sci.* 16. doi: 10.3389/fpls.2025.1506524

Lu, F., Shangguan, H., Yuan, Y., Yan, Z., Yuan, T., Yang, Y., et al. (2025). LeafConvNeXt: Enhancing plant disease classification for the future of unmanned farming. *Comput. Electron. Agric.* 233, 110165. doi: 10.1016/j.compag.2025.110165

Mahmud, B. U., Al Mamun, A., Hossen, M. J., Hong, G. Y., and Jahan, B. (2024). Light-weight deep learning model for accelerating the classification of mango-leaf disease. *Emerging Sci. J.* 8, 28–42. doi: 10.28991/ESJ-2024-08-01-03

Mia, M. R., Roy, S., Das, S. K., and Rahman, M. A. (2020). Mango leaf disease recognition using neural network and support vector machine. *Iran J. Comput. Sci.* 3, 185–193. doi: 10.1007/s42044-020-00057-z

Nirob, M. A. S., Bishshash, P., Siam, A.K.M.F.K., Mia, S., Khatun, T., and Uddin, M. S. (2024). *Mango dataset: A comprehensive resource for agricultural research and disease detection* (Daffodil International University, Bangladesh: Mendeley Data). doi: 10.17632/fn8dgf4hb5.1

Padshetty, S., and Umashetty, A. (2024). Agricultural innovation through deep learning: A hybrid CNN-Transformer architecture for crop disease classification. *J. Spatial Sci.*, 1–32. doi: 10.1080/14498596.2024.2355225

Pahati, M., De Jesus, L. C., Reyes, R., Villarroel, J. M., and Angeles, M. R. (2025). "Detecting mango leaf diseases using google teachable machine for sustainable agriculture," in *2025 international conference on artificial intelligence in information and communication (ICAIIC)* (Fukuoka, JapanFukuoka, Japan: IEEE), 0611–0614.

Patel, R. K., Chaudhary, A., Chouhan, S. S., and Pandey, K. K. (2024). Mango leaf disease diagnosis using Total Variation Filter Based Variational Mode Decomposition. *Comput. Electrical Eng.* 120, 109795. doi: 10.1016/j.compeleceng.2024.109795

Pratap, V. K., and Kumar, N. S. (2024). Deep learning based mango leaf disease detection for classifying and evaluating mango leaf diseases. *Fusion: Pract. Appl.* 15, 261–77. doi: 10.54216/FPA.150222

Puranik, S. S., Hanamakkanavar, S. R., Bidargaddi, A. P., Ballur, V. V., Joshi, P. T., SM, M., et al. (2024). "MobileNetV3 for mango leaf disease detection: an efficient deep learning approach for precision agriculture," in *2024 5th international conference for emerging technology (INCET)* (Belgaum, India: IEEE), 1–7.

Rahman, M. S., Hasan, R., and Mojumdar, M. U. (2024). *Mango leaf disease dataset* (Daffodil International University, Bangladesh: Mendeley Data). doi: 10.17632/7ghdbftp54.1

Rao, U. S., Swathi, R., Sanjana, V., Arpitha, L., Chandrasekhar, K., and Naik, P. K. (2021). Deep learning precision farming: Grapes and mango leaf disease detection by transfer learning. *Global Transitions Proc.* 2, 535–544. doi: 10.1016/j.gltp.2021.08.002

Rozenfeld, S., Kalo, N., Naor, A., Dag, A., Edan, Y., and Alchanatis, V. (2024). Thermal imaging for identification of malfunctions in subsurface drip irrigation in orchards. *Precis. Agric.* 25, 1038–1066. doi: 10.1007/s11119-023-10104-x

Saleem, R., Shah, J. H., Sharif, M., and Ansari, G. J. (2021b). Mango leaf disease identification using fully resolution convolutional network. *Computers Materials Continua* 69, 3581–3601. doi: 10.32604/cmc.2021.017700

Saleem, R., Shah, J. H., Sharif, M., Yasmin, M., Yong, H. S., and Cha, J. (2021a). Mango leaf disease recognition and classification using novel segmentation and vein pattern technique. *Appl. Sci.* 11, 11901. doi: 10.3390/app112411901

Shakib, M. M. H., Mustofa, S., and Ahad, M. T. (2024). *MLD24: an image dataset for mango leaf disease detection* (Daffodil International University, Bangladesh: Mendeley Data). doi: 10.17632/6dvpywm2m2.1

Shehu, H. A., Ackley, A., Mark, M., Eteng, O. E., Sharif, M. H., and Kusetogullari, H. (2025). YOLO for early detection and management of Tuta absoluta-induced tomato leaf diseases. *Front. Plant Sci.* 16. doi: 10.3389/fpls.2025.1524630

Singh, Y. P., Chaurasia, B. K., and Shukla, M. M. (2024). Deep transfer learning driven model for mango leaf disease detection. *Int. J. System Assur. Eng. Manage.* 15, 4779–4805. doi: 10.1007/s13198-024-02480-y

Tao, Y., Chang, F., Huang, Y., Ma, L., Xie, L., Su, H., et al. (2022). Cotton disease detection based on ConvNeXt and attention mechanisms. *IEEE Journal of Radio Frequency Identification*, 6, 805–809. doi: 10.1109/JRFID.2022.3206841

Varma, T., Mate, P., Azeem, N. A., Sharma, S., and Singh, B. (2025). Automatic mango leaf disease detection using different transfer learning models. *Multimedia Tools Appl.* 84, 9185–9218. doi: 10.1007/s11042-024-19265-x

Wei, C., Shan, Y., and Zhen, M. (2025). Deep learning-based anomaly detection for precision field crop protection. *Front. Plant Sci.* 16. doi: 10.3389/fpls.2025.1576756

Yavuz, E., and Aydemir, Ö. (2016). "Olfaction recognition by EEG analysis using wavelet transform features," in *2016 international symposium on INnovations in intelligent sysTems and applications (INISTA)* (Sinaia, Romania: IEEE), 1–4.

Zhang, B., Wang, Z., Ye, C., Zhang, H., Lou, K., and Fu, W. (2025). Classification of infection grade for anthracnose in mango leaves under complex background based on CBAM-DBIRNet. *Expert Syst. With Appl.* 260, 125343. doi: 10.1016/j.eswa.2024.125343

Zhou, Y., Wang, X., Zhang, M., Zhu, J., Zheng, R., and Wu, Q. (2019). MPCE: A maximum probability based cross entropy loss function for neural network classification. *IEEE Access* 7, 146331–146341. doi: 10.1109/ACCESS.2019.2946264