scientific reports

Check for updates

OPEN

# Dual-modality fusion for mango disease classification using dynamic attention based ensemble of leaf & fruit images

Muhammad Mohsin[1], Muhammad Shadab Alam Hashmi[1]✉, Irene Delgado Noya[2,3,4,5], Helena Garay[2,6,7], Nagwan Abdel Samee[8] & Imran Ashraf[9]✉

**Mango is one of the most beloved fruits and plays an indispensable role in the agricultural economies of many tropical countries like Pakistan, India, and other Southeast Asian countries. Similar to other fruits, mango cultivation is also threatened by various diseases, including Anthracnose and Red Rust. Although farmers try to mitigate such situations on time, early and accurate detection of mango diseases remains challenging due to multiple factors, such as limited understanding of disease diversity, similarity in symptoms, and frequent misclassification. To avoid such instances, this study proposes a multimodal deep learning framework that leverages both leaf and fruit images to improve classification performance and generalization. Individual CNN-based pre-trained models, including ResNet-50, MobileNetV2, EfficientNet-B0, and ConvNeXt, were trained separately on curated datasets of mango leaf and fruit diseases. A novel Modality Attention Fusion (MAF) mechanism was introduced to dynamically weight and combine predictions from both modalities based on their discriminative strength, as some diseases are more prominent on leaves than on fruits, and vice versa. To address overfitting and improve generalization, a class-aware augmentation pipeline was integrated, which performs augmentation according to the specific characteristics of each class. The proposed attention-based fusion strategy significantly outperformed individual models and static fusion approaches, achieving a test accuracy of 99.08%, an F1 score of 99.03%, and a perfect ROC-AUC of 99.96% using EfficientNet-B0 as the base. To evaluate the model's real-world applicability, an interactive web application was developed using the Django framework and evaluated through out-of-distribution (OOD) testing on diverse mango samples collected from public sources. These findings underline the importance of combining visual cues from multiple organs of plants and adapting model attention to contextual features for real-world agricultural diagnostics.**

**Keywords** Plant disease detection, Multimodal approach, Class-aware augmentation, Modality attention fusion, Out-of-distribution

Mango, scientifically known as Mangifera indica and celebrated as the "King of Fruits," holds immense cultural, nutritional, and economic significance, particularly across South Asia and tropical regions. It is one of the world's top-harvested fruits, which originated in South Asia with over 4,000 years of cultivation and now comprises more than 500 cultivars worldwide[1]. In 2019, the global production of mango reached approximately 51 million tons, with India alone accounting for about 55% of the total harvest, which was approximately 24 million tons in 2020[2]. Also, Pakistan stands as a major producer of mango; however, it exports only 5 to 7 percent of its total mango output[3].

[1]Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan. [2]Universidad Europea del Atlantico, Isabel Torres 21, 39011 Santander, Spain. [3]Universidad Internacional Iberoamericana, 24560 Campeche, Mexico. [4]Fundacion Universitaria Internacional de Colombia, Bogota, Colombia. [5]Universidad de La Romana, La Romana, República Dominicana. [6]Universidad Internacional Iberoamericana Arecibo, Puerto Rico 00613, USA. [7]Universidade Internacional do Cuanza, Cuito, Bie, Angola. [8]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia. [9]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea. ✉email: shadab.alam@kfueit.edu.pk; imranashraf92@gmail.com

Asians, specifically Southeast Asians, are not only obsessed with mango for its flavor but also yearn for it due to its rich nutritional profile. The pulp contains high levels of carbohydrates, mainly glucose, fructose, and sucrose, which make up 15-18% w/v of its weight. It is also a great source of Vitamin C, with unripe cultivars presenting as much as 88mg per 100g, while at maturity, pulp levels range from 21 to 51 μg/g in peel extracts. And not just to extinguish the temptation, mangoes also have several bioactive compounds, including phenolics, flavonoids, and carotenoids, that lend antioxidant, anti-inflammatory, and antimicrobial properties[4].

Similar to other fruits, mango cultivation is also threatened by different diseases, which can drastically compromise mango production and farmer livelihoods. For instance, anthracnose, one of the lethal diseases we diagnosed in this study, is caused by Colletotrichum gloeosporioides and causes losses of up to 39% in some regions. Other diseases that mangoes have to face include powdery mildew, bacterial canker, black rot, dieback, sooty mold, and gall midge infestations. A multi-year study in humid tropical regions found that unprotected orchards suffered 47% average yield losses due primarily to diseases like powdery mildew, anthracnose, and stem-end rot[5]. Moreover, mango decline, characterized by dieback, root rot, and bacterial leaf spot, is prevalent across key regions like Rahim Yar Khan and Multan, affecting 90–100% of orchards and reducing production capacity by 37–53% disease severity indices[6].

Such a diversity of diseases makes it almost impossible to be diagnosed by any human, and to mitigate this challenge, many studies have employed deep learning on mango leaf images. For instance, a CNN achieved 99% accuracy across eight leaf disease categories in a mobile-based system[7]. Another model using attention-based multi-scale feature fusion on the MangoLeafBD dataset obtained 99% accuracy[8]. However, limitations still prevail due to multiple reasons, such as imbalanced datasets, symptoms of different diseases appearing visually similar, and most models being trained on clean, lab-grade images rather than field-acquired data. Most of this research is unimodal because it focuses exclusively either on leaves or fruits, yet real infections can affect both leaves and fruits, and integrating both could improve detection. The main contributions of this study are listed below.

- *Dual-model attention-based fusion architecture:* We propose a dual-stream convolutional framework comprising two specialized models: one trained on mango leaf images and the other on fruit images. To leverage complementary visual cues from both modalities, we implement a dynamic attention-based late fusion strategy that adaptively weights their predictions. This improves robustness in cases where one modality may be degraded or ambiguous. As a direction for future work, we suggest exploring cross-modality alignment through contrastive learning to project both inputs into a shared embedding space.
- *Advanced and class-aware augmentation pipeline:* Rather than using generic augmentations (e.g., random flips or color jitter), we design a class-conditional augmentation strategy that mimics real-world distortions likely to occur in mango orchards. For instance, Anthracnose samples are augmented using random rotations and leaf angle shifts, and Bacterial Black Spot images avoid rotation-based transformations to preserve lesion consistency; however, their brightness was increased in the Fruit dataset.

After conducting a thorough literature review, it was found that dual modality fusion models are not frequently applied, specifically in the case of disease detection in mango plants. It is identified as an important research gap, since in real-world scenarios, unimodal classification often fails in cases where one modality (leaf or fruit) alone is insufficient. Farmers frequently face such challenges, and the proposed approach aims to provide an alternative solution by leveraging complementary information from both modalities. This work, therefore, holds importance for plant disease classification and in other domains too.

The second contribution is the use of class-aware augmentation. Despite not being very fruitful in achieving substantial performance improvements in this study, it still demonstrates the potential of exploring class-specific strategies rather than relying solely on conventional augmentation methods. We believe this adds exploratory value and can encourage future work in this direction.

## Literature review

There have been several research studies in the domain of plant disease classification, involving grapes[9], tomato[10], apple[11,12], etc. This section covers some of the most recent and advanced literature in the realm of traditional deep learning methods and multi-modal approaches for mango disease detection.

To enhance accuracy,[13] introduces a Plant-CNN-ViT ensemble that capitalizes on the complementary strengths of four pre-trained models. These models include Vision Transformer (ViT), ResNet-50, DenseNet-201, and Xception. Here, the ViT was specifically dedicated to capturing dependencies and important features, while ResNet-50 introduced residual connections. Similarly, DenseNet-201 ensured dense connections to capture intricate leaf patterns. Finally, Xception used separable convolutions that drastically reduced computational cost. This methodology achieved a remarkable accuracy of 100% on the Flavia, Folio, and Swedish Leaf datasets, and 99.83% on the MalayaKew Leaf dataset.

Another research in the domain of deep learning proposed MobilePlantViT, which is a hybrid vision transformer architecture that optimizes resource efficiency while maintaining high performance[14]. This architecture extracts essential features via a stack of grouped convolutions that are fused with convolutional attention modules. It was then evaluated on both standardized datasets (PlantVillage, Coconut) and real-world datasets (CCMT: Cashew, Cassava, Maize, Tomato, and Sugarcane). This scored an accuracy of 99.57% on PlantVillage and 99.20% on Coconut. Among the CCMT datasets, MobilePlantViT demonstrated strong generalization capabilities, achieving 95.04% on CCMT-Cashew, 94.34% on CCMT-Cassava, and 92.76% on Sugarcane. Although performance was relatively lower for CCMT-Maize (81.05%) and CCMT-Tomato (80.05%), macro and weighted F1-scores remained above 0.80, indicating robust multi-class classification under diverse field conditions. The study further showed that domain-specific pretraining (on PlantVillage) significantly

| Study | Modality | Method | Datasets | Results |
|---|---|---|---|---|
| [13] | Only leaf | Plant-CNN-ViT- CNN + ViT Ensemble (ResNet50, DenseNet201, Xception, ViT) | Flavia, Folio, Swedish, MalayaKew | 100%, 99.83% |
| [14] | Only leaf | MobilePlantViT- Lightweight ViT + grouped convs + attention | PlantVillage, Coconut, CCMT (Cashew, Cassava, etc.) | 99.57% (PlantVillage)95.04% (Cashew) |
| [15] | Only leaf | CNN (VGG16, InceptionV3, DenseNet201) + ViT | Apple, Corn (4-class each) | 99.24%, 98% |
| [16] | Visually MultiModal ( RGB + segmented leaf image) | Multi-head DenseNet | PlantVillage (38 classes) | 98.17% F1 |
| [17] | Multimodal (Image & sensor data) | ResNet-50 + Attention Fusion | Custom Eggplant Dataset (5 classes) | 92% Acc, 94% Prec, 90% Recall |

**Table 1**. Summary of recent vision-based and multimodal plant disease detection studies.

| Dataset | Number of images |
|---|---|
| MLD24 | 2,353 |
| SenMangoFruit DDS | 862 |
| Mango fruit early disease detection (Kaggle) | 400 |
| Damage cause mango project (Roboflow) | 609 |

**Table 2**. Summary of mango image datasets used in this study.

improved performance for smaller datasets like Maize and Tomato, with up to a 4.5% accuracy gain. Likewise, another study used a similar method by creating a hybrid framework[15]. In this approach, an ensemble of popular CNN models, VGG16, Inception V3, and DenseNet201, was used to extract overall (global) features. Then, a ViT model was added to focus on fine (local) details to improve the accuracy of plant disease detection. The combined model was tested on two public datasets, one for apples and one for corn, each with four types of diseases. This hybrid model also achieved remarkable accuracy, with 99.24% on the Apple dataset and 98% on the Corn dataset.

In the domain of multimodal and adaptive fusion, this research developed a multi-headed DenseNet-based architecture that considers two images as input[16]. It fuses RGB and segmented images from the publicly available PlantVillage dataset, which consists of 54,183 images across 38 classes. A five-fold cross-validation technique was employed, achieving an average accuracy of 98.17%, recall of 98.17%, precision of 98.16%, and an F1 score of 98.12%, which shows that the proposed model can precisely differentiate various plant diseases with different characteristics by image fusion.

The research[17] proposed a multimodal and attention mechanism model for eggplant disease detection, which integrates images and sensor data to improve the model's ability. Images were explicitly collected to create a diverse dataset of five diseases for this research, where the image input undergoes preprocessing before being passed through ResNet-50 to extract high-level features. After processing through the respective networks, the data enters the multimodal fusion stage, where the extracted features are combined using a weighted fusion strategy that proportionally adjusts the influence of image and sensor modalities. As a result, the model achieved a precision of 94%, a recall of 90%, and an accuracy of 92%. Table 1 provides a summary of the discussed works.

## Methodology

### Dataset preparation

This section covers all the steps that were executed during this research to ensure the data fed to the models was appropriate, because the success of any deep learning-based image classification model relies heavily on the quality, quantity, and variability of the dataset used during training. In the context of mango disease classification, which is the domain of this research, preparing a reliable dataset is especially challenging due to multiple hindrances such as intra-class variability, inter-class similarity, and class imbalance. We tried to mitigate these issues with modern state-of-the-art methodologies that are mentioned below in chronological order.

### Dataset acquisition

During this research, the major issue was the limited data availability for the same classes in both types of datasets, mango leaf and mango fruit. Since we were performing late fusion, which required having the same classes in both types of datasets, we had to use a total of four (4) datasets to fulfill that requirement. The motivation for this dual-dataset approach stems from the observation that certain mango diseases manifest more prominently on leaves, while others are more visible on fruits. By leveraging both modalities, we aim to improve diagnostic coverage and accuracy in real-world scenarios. Table 2 contains the number of images acquired from each dataset.

*MLD24: an image dataset for mango leaf disease detection*

The first dataset we used was downloaded from Mendeley Data, which is a free and secure cloud-based communal repository for storing data. This dataset contains 2,336 processed images, which we used because augmentation was later implemented only on the training set after splitting[18]. It contains eight classes, collected

from mango trees in Bangladesh, using a mobile phone camera over a span of 20 days. The Bacterial Canker class was renamed as Bacterial Black Spot because, based on agronomic literature and expert consultation, we observed that the visual symptoms represented in this class, namely small black lesions with yellow halos, often expanding into necrotic spots, more accurately align with what is commonly referred to as "Bacterial Black Spot" in mango pathology. The term "Bacterial Canker" is sometimes used interchangeably in regional contexts, but in mango disease taxonomy, Bacterial Black Spot (BBS), caused by Xanthomonas campestris pv. Mangifera indica is the more precise and internationally accepted nomenclature[19]. Therefore, for our needs, clarity and consistency with global literature, we renamed the class "Bacterial Canker" to "Bacterial Black Spot" in this study.

*SenMangoFruitDDS*

This dataset contains images of mango fruit, spanning across four disease classes, which include Alternaria, Anthracnose, Black Mould Rot, and Stem-End Rot. An additional class in this dataset was that of healthy mangoes, and this dataset was also downloaded from Mendeley Data. The images were in 224×224 size in JPG format, with two versions, one with the original images and the second with the background removed. We considered the version with the background removed[20].

*Mango fruit early diseases detection*

Kaggle is a haven for data scientists because it contains hundreds of datasets to train ML models. We downloaded this dataset from Kaggle, and it was the very first one we considered, but it contained only 400 images evenly distributed across four categories: 'Anthracnose', 'Bacterial Black Spot', 'Healthy', and 'Multiple'. The 'Multiple' class shows images that contain any combination of more than one disease. That is the reason why we were seeking these four classes in other datasets as well. Also, the images were of low quality, which later went through augmentation to make them more robust[21].

*Damage cause mango project*

Roboflow is an online web app that provides computer vision tools to build and deploy models, and we used the data from one such project to extract a single class from it. The entire dataset contained 3,225 images annotated in COCO format, and we extracted only the Bacterial Black Spot class from it because it was the only one we were lacking. This was done using COCO JSON annotations with Python libraries such as json and collections. The images were extremely raw and taken at unsymmetrical angles, so we had to remove their backgrounds before merging them. This was done using rembg, which leverages a powerful deep learning model named U-2-Net for foreground-background separation. Later, the background was changed to white, and images were manually filtered to completely remove inconsistent or faulty samples. Figure 1 illustrates how Python libraries were used to remove the background and later converted into solid white.

This study uses four public datasets with different data acquisition protocols, e.g., resolution, presence/absence of background, and growth stages. Since each dataset was acquired from a different source, it provided us with the advantage of diversity, which is a strong point in the domain of machine learning. However, it also introduced certain challenges. For example, the mango fruit images obtained from RoboFlow contained random backgrounds such as trees, mud, and sky, which severely misled the models during training. To mitigate this issue, we opted to remove the backgrounds, allowing the model to focus solely on the target. This significantly improved the model's learning performance, as the removed backgrounds were replaced with plain white while keeping the image dimensions intact.

In other datasets, differences in lighting were present, with some images captured indoors and others outdoors. We left these variations untouched for two reasons. The first was that class-aware augmentation, applied during preprocessing, is adopted to standardize each class. The second, and more important reason is that variations in saturation, contrast, or simply 'lighting,' helped the model generalize better across multiple environmental conditions.

(a) (b)

**Fig. 1**. Sample image showing background removal for the bacterial black spot class, (**a**) Original image, and (**b**) After background is removed.

## Class-level modality pairing

Before splitting, the datasets mentioned above that were collected independently were merged and aligned according to the classes to enable meaningful multimodal fusion. Here, we ensured that rather than aligning based on one-to-one image correspondences, which are often impractical in real-world agricultural datasets, we utilized the class labels as the unit of alignment. In other words, for each class, for instance, Anthracnose or Bacterial Black Spot, we ensured a one-to-one semantic mapping between the corresponding leaf and fruit images. This class-level modality pairing allows both models, one trained on leaf images and the other on fruit images, to learn complementary features for the same disease category, even though the individual images and their quantity differ.

This setup makes the late fusion strategy we used here semantically consistent because the predictions from the leaf model and the fruit model both refer to the same disease class, even if the image inputs were different. This not only helped us preserve class coherence across modalities but also enhanced the robustness of the final predictions by incorporating disease cues from two anatomically distinct parts of the mango plant.

## Dataset splitting

After ensuring that we had enough images to train the models by merging multiple datasets, we split the data into three distinct subsets by allocating 70% of the images for training, 15% for validation, and 15% for testing. This was done to ensure robust model evaluation and to prevent overfitting in both the leaf and fruit datasets. The training set was used to fit the model parameters, while the validation set was used for early stopping and model selection. The final performance metrics were calculated on the independent test set to simulate real-world deployment conditions. This 70/15/15 split provides a good trade-off between model training depth and evaluation reliability. It follows best practices in machine learning literature and has been commonly adopted in similar plant disease classification tasks[22,23]. Table 3 summarizes the number of images per class in the training, validation, and test sets for both the leaf and fruit datasets.

## Class-based augmentation

Prior research utilized uniform augmentation pipelines where they performed the same augmentation on each class[24]. However, in contrast to this generic strategy, we introduced a class-aware augmentation strategy to reflect the real-world variability in mango disease presentation. This variability exists not just across classes but also across datasets, as each class and each type of dataset (leaf and fruit) exhibits distinct visual characteristics and tolerance to distortion. Therefore, applying uniform augmentation can distort or compromise class-specific features. Class-aware augmentation, in layman's terms, is an image enhancement strategy where different classes are augmented with transformations according to their specific characteristics, rather than applying the same augmentations uniformly across all classes.

For instance, leaf diseases like Powdery Mildew or Sooty Mold, used in the Multiple category in this research and characterized by fine-textured overlays or superficial surface growth, are highly sensitive to spatial and contrast-based transformations. To preserve their delicate visual patterns, we avoided strong blur-based augmentations and instead applied only minor geometric shifts and subtle brightness adjustments. In contrast, diseases such as Anthracnose, which often produce large, irregular necrotic patches with dark pigmentation, were well-suited to more aggressive augmentations. For this class, we introduced motion blur to simulate camera shake and varying capture angles typically encountered in orchard conditions. Similarly, Bacterial Black Spot (BBS), characterized by small, angular, greasy lesions with yellow halos, was found to benefit from Gaussian blur, which helped model variability introduced by focus issues or environmental interference during image capture. Figure 2 shows the original and augmented images for the Leaf dataset.

By assigning augmentation types based on the structural characteristics of each disease class, we ensured that the learning process emphasized disease-relevant features while improving model robustness against real-world distortions. It is important to note that these augmentations were applied only to the training set, while the validation and testing sets were kept raw. Table 4 represents the complete types and variations of augmentation applied to different classes, and Fig. 3 shows the results of augmentation on the Mango Fruit dataset. Also, in most cases, data augmentation is employed to increase the size of the training set by generating additional samples. However, in this work, we did not apply oversampling-based augmentation; instead, we utilized

| Category | Anthracnose | Bacterial BS | Healthy | Multiple | Total |
|---|---|---|---|---|---|
| Leaf dataset | | | | | |
| Train | 228 | 251 | 175 | 991 | 1645 |
| Validation | 48 | 54 | 37 | 212 | 351 |
| Test | 50 | 55 | 38 | 214 | 357 |
| Total | 326 | 360 | 250 | 1,417 | 2,353 |
| Fruit dataset | | | | | |
| Train | 162 | 496 | 215 | 435 | 1308 |
| Validation | 34 | 106 | 46 | 93 | 279 |
| Test | 36 | 107 | 47 | 94 | 284 |
| Total | 232 | 709 | 308 | 622 | 1,871 |

**Table 3**. Image distribution across the leaf and fruit datasets.

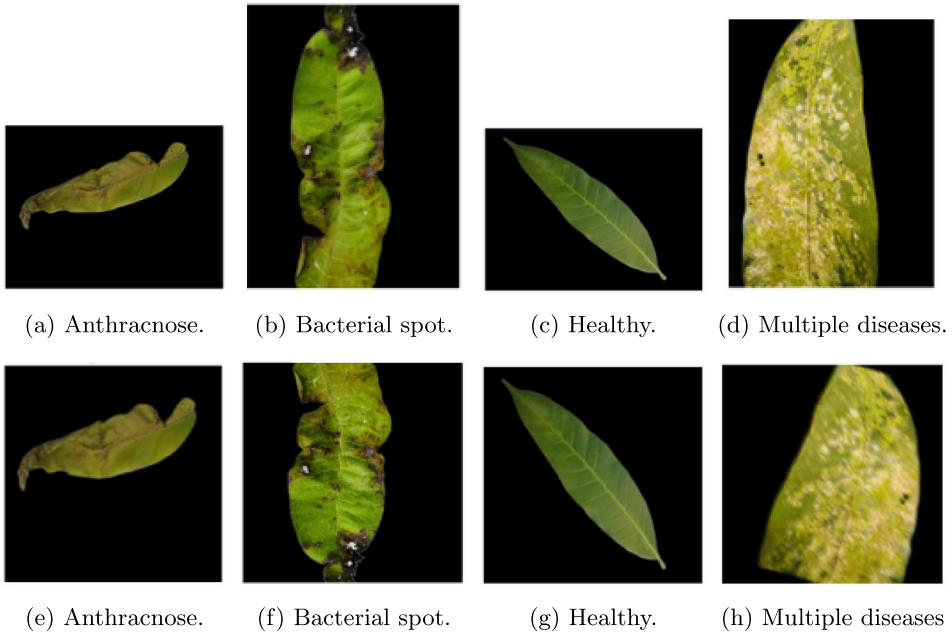



**Fig. 2**. Sample image showing original images (**a**–**d**), and class-based augmentation applied on the Leaf dataset (**e**–**h**) with respect to their classes.

| Class | Dataset | Augmentations applied |
|---|---|---|
| Anthracnose | Leaf | Resize (224×224), Horizontal Flip ($p = 0.5$), Rotate (±25°, $p = 0.8$), Motion Blur (limit=3, $p = 0.2$) |
| | Fruit | Resize (224×224), Horizontal Flip ($p = 0.5$), Motion Blur ($p = 0.4$), Rotate (±15°, $p = 0.5$) |
| Bacterial black spot | Leaf | Resize (224×224), Vertical Flip ($p = 0.5$), ShiftScaleRotate (±5% shift/scale, ±10° rotate, $p = 0.7$) |
| | Fruit | Resize (224×224), Vertical Flip ($p = 0.5$), Gaussian Blur ($p = 0.4$), Rotate (±10°, $p = 0.4$) |
| Healthy | Leaf | Resize (224×224), Rotate (±10°, $p = 0.5$), Random Brightness/Contrast ($p = 0.1$) |
| | Fruit | Resize (224×224), Horizontal Flip ($p = 0.3$), Vertical Flip ($p = 0.3$), Rotate (±5°, $p = 0.3$) |
| Multiple | Leaf | Resize (224×224), Horizontal Flip ($p = 0.5$), Vertical Flip ($p = 0.5$), Rotate (±30°, $p = 0.8$), Motion Blur ($p = 0.3$) |
| | Fruit | Resize (224×224), Horizontal Flip ($p = 0.5$), Vertical Flip ($p = 0.5$), Rotate (±20°, $p = 0.5$) |

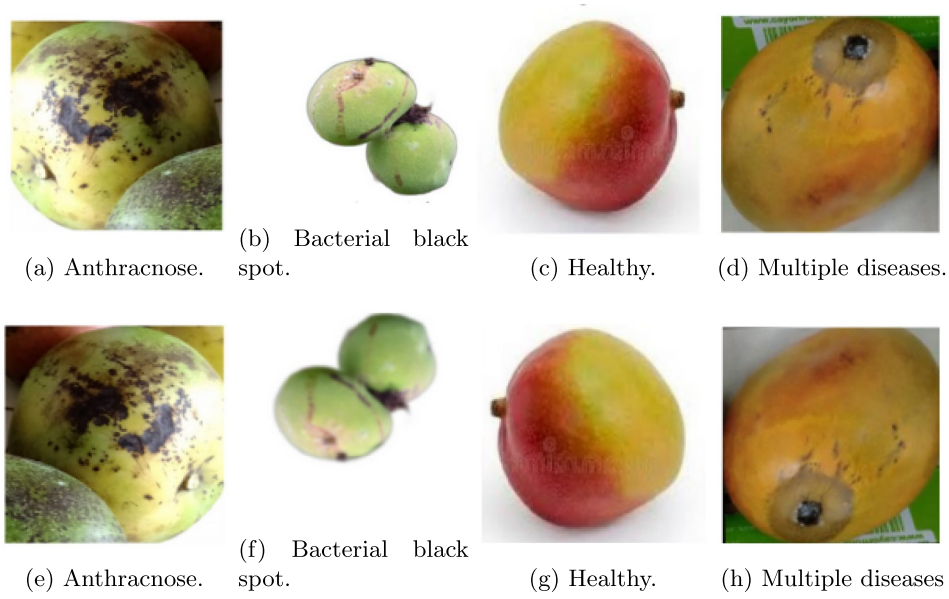
**Table 4**. Class-specific augmentations applied to mango leaf and fruit datasets.

transformation augmentation, also referred to as on-the-fly augmentation, where stochastic transformations are applied to existing images during training. This strategy preserves the original dataset size while introducing variability at each epoch, thereby enhancing model generalization without artificially inflating the dataset.

## Proposed approach

After overcoming the challenge of data limitation, we employed a robust pipeline to ensure reliable mango disease classification. We begin by detailing the dataset preparation process, followed by the training of individual CNN models on leaf and fruit images separately. Subsequently, we present a novel modality fusion approach, which integrates features from both modalities to enhance classification performance. We also describe the implementation of cross-validation strategies to ensure robustness and generalization. Finally, we evaluate the practical applicability of the models through real-world out-of-distribution testing using an interactive prediction app. Figure 4 demonstrates the holistic workflow pipeline implemented in this research.

In order to design and test a robust multimodal classification pipeline for mango disease detection, we utilized five widely recognized CNN architectures, which include ResNet18, ResNet50, MobileNetV2, EfficientNet-B0, and ConvNeXt-Tiny. Each of these models was pretrained on the ImageNet dataset, which is a large-scale database consisting of over 14 million high-resolution images categorized using WordNet synsets, and was fine-tuned on the mango leaf and fruit datasets independently using transfer learning techniques by modifying their last layers with respect to the classes. Also, CNN models built from scratch require hundreds of images and consume significant resources and time to train. That is why we opted to use pre-trained models.



(a) Anthracnose. (b) Bacterial black spot. (c) Healthy. (d) Multiple diseases.

(e) Anthracnose. (f) Bacterial black spot. (g) Healthy. (h) Multiple diseases

**Fig. 3**. Sample image showing original images (**a**–**d**), and class-based augmentation (**e**–**h**), applied on the Fruit dataset with respect to their classes.

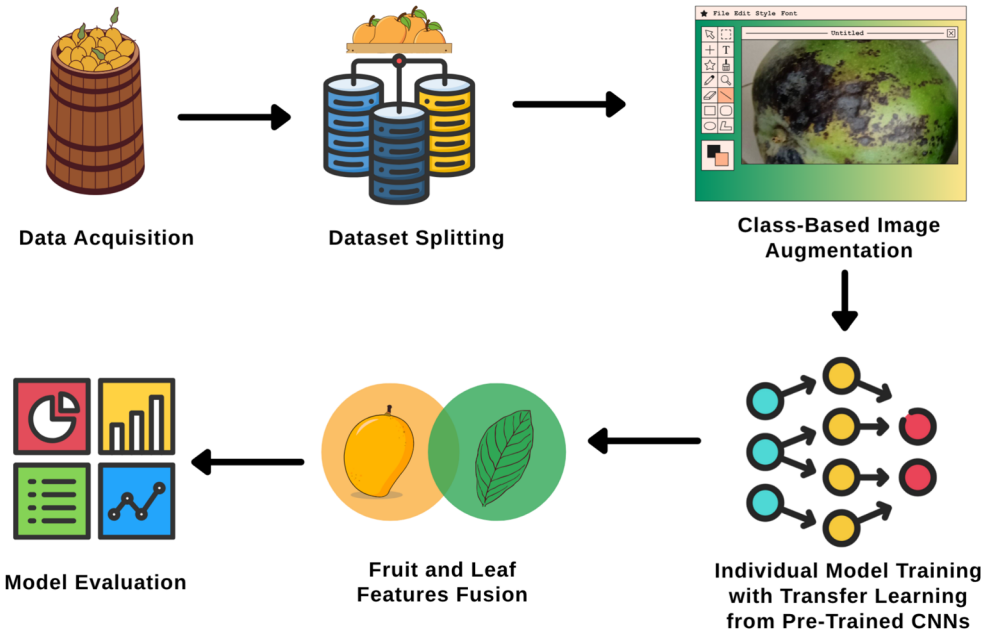



**Fig. 4**. Holistic experimental workflow diagram of the proposed approach.

*ResNet-50*

ResNet-50 (Residual Network) addresses the vanishing gradient problem in deep networks, whose goal is to minimize the loss by introducing residual connections that allow gradients to flow through identity shortcuts. In each residual block, the output is computed as:

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{1}$$

where $\mathcal{F}$ represents the residual mapping learned by a stack of layers, and x is the input that is directly passed through a shortcut connection. This formulation allows the network to refine features rather than relearn them entirely, facilitating better convergence and representational learning.

ResNet-50 comprises 50 layers and utilizes bottleneck blocks, which stack 1×1, 3×3, and 1×1 convolutions to reduce computational complexity while preserving depth. It has become a widely adopted architecture in image classification and transfer learning tasks due to its balance between depth and efficiency[11,25].

*MobileNetV2*
MobileNetV2 is a lightweight model specifically designed for mobile and embedded vision applications. It uses depthwise separable convolutions, which reduce computation, along with inverted residuals and linear bottlenecks that preserve low-dimensional features while enhancing learning efficiency. This means that instead of reducing channel dimensions, the block expands the input, applies depthwise convolution, and then compresses it back linearly without activation. This preserves information while allowing non-linear transformations in the middle layers[26].

$$y = P(D(x)) \tag{2}$$

where $D$ is depthwise convolution and $P$ is pointwise convolution.

*EfficientNet-B0*
EfficientNet is a well-known pre-trained model that leverages a smart scaling method to adjust the model's depth, width, and input resolution[27]. The base model, EfficientNet-B0, is built using MBConv blocks, which are improved using squeeze-and-excitation (SE) modules. The SE modules support the network to focus more on important features by adjusting how much attention is given to each one. It achieves a remarkable accuracy vs. parameter efficiency trade-off, which makes it highly effective for training on small-to-medium-sized datasets such as mango leaves and fruits[28].

*ConvNeXt-tiny*
ConvNeXt is a modern convolutional architecture inspired by the success of Vision Transformers (ViTs) but optimized for convolutional operations. It applies insights from transformer design, such as normalization, GELU activations, and larger kernel sizes, to a purely convolutional framework. Some key innovations in ConvNeXt include the replacement of traditional ReLU with GELU for smoother activation and the substitution of BatchNorm with LayerNorm in certain blocks. It uses depthwise convolutions and inverted bottlenecks, similar to those in MobileNet and EfficientNet. In this case, ConvNeXt-Tiny serves as a powerful backbone for leaf and fruit classification with strong generalization[29].

## Novel modality attention fusion

In practical use or real-world settings, one modality may be more informative than the other. For instance, Anthracnose symptoms may appear more prominently on fruits, while Bacterial Black Spot may show clearer patterns on leaves. In such circumstances, a fixed fusion weight can't capture sample-specific or class-specific variability. That is why we introduce an adaptive fusion mechanism.

The modality attention fusion is a framework that learns to dynamically assign weights to different input types (e.g., leaf and fruit images), so the model can focus more on the modality that provides stronger cues for disease identification. Modality fusion offers several advantages.

- Mango diseases differ in their inception. For example, powdery mildew and anthracnose first appear on leaves. On the other hand, fruit rot and stem-end rot become more visible on mango fruits. Combining both images offers a more complete diagnostic solution.
- Images from multiple modalities often reduce false negatives due to the model's learning from multiple perspectives. It improves the overall performance of the model.
- Model when training on leaf and fruit images can become more generalizable to real-world conditions, particularly where the disease appearance varies across various parts of the plant.
- Leaf images are good for early detection, while fruit images support disease detection for later stages of growth. Combining these images can provide an end-to-end crop protection solution.
- Using both images for model training offers a more robust solution. The farmers can use any of these organs, whichever shows the disease symptoms first. In addition, multi-modal training makes the model robust for variable lighting and environmental conditions.

Unlike simple averaging or static weight fusion, the approach introduces a learnable attention mechanism that adjusts the fusion weights based on the semantic reliability of each modality's prediction[30].

Let $z_L \in R^C$ be the logits from the leaf model, $z_F \in R^C$ be the logits from the fruit model, and $C$ be the number of classes (e.g., 8). We first pass these logits through the softmax function to obtain class probability vectors.

$$p_L = \text{softmax}(z_L), \quad p_F = \text{softmax}(z_F) \tag{3}$$

Then, we compute the attention weight for each modality:

$$\alpha = \sigma(w^\top [z_L \,||\, z_F] + b) \tag{4}$$

where,

- $[z_L \,||\, z_F] \in R^{2C}$ is the concatenated logits vector,

- $w \in R^{2C}$ and $b \in R$ are learnable parameters,
- $\sigma(\cdot)$ is the sigmoid function, resulting in $\alpha \in (0, 1)$.

The final fused prediction is computed as:

$$p_{\text{fused}} = \alpha \cdot p_L + (1 - \alpha) \cdot p_F \tag{5}$$

This formulation allows the model to dynamically emphasize one modality over the other based on the sample characteristics, which enhances generalization and robustness, as shown in Fig. 5.

## Django-based prediction web app

In order to test real-world usability and model interpretability, we developed a fully functional full-stack web app using the Django framework and Python for the back end, and HTML, CSS, JavaScript, and Tailwind CSS for the front end. This application integrates the complete pipeline, allowing end users to test all four trained convolutional network models with three prediction modes. The first is leaf-only classification, the second is fruit-only classification, and the third is fusion-based classification using the proposed Modality Attention Fusion mechanism.

In this app, users can upload a single image for leaf or fruit-only classification, or paired leaf and fruit images for fusion, to receive real-time disease predictions with their respective class labels and confidence scores. Figure 6 is a screenshot of the user interface with a practical example using completely new images, which were not used during training, where the fused model prediction is displayed along with its respective confidence.

This web application was deployed on Vercel, which is a cloud-based platform optimized for front-end frameworks and static site generation. However, for the model hosting and inference, we utilized Hugging Face Spaces, which provides a scalable infrastructure for deploying machine learning models via RESTful APIs. The system architecture is modular, in which the front-end is developed in Django, acting as the client interface. It communicates asynchronously with the Hugging Face back-end through HTTP POST requests (utilizing FastAPI), sending user inputs for prediction. The back-end server dynamically loads the pre-trained models, performs inference, and returns the predicted disease class to the front end in real time. This client-server architecture ensures low latency, high modularity, and easy scalability. The live application can be accessed at the following link: *Mango Disease Classifier*. For real-world use, this app can be accessed from a browser or mobile device, as it is mobile-responsive. Images can be uploaded to get classified according to the type of disease they carry.

## Training configuration

All the pretrained models used in this study were independently trained on the leaf and fruit datasets. During each epoch, the model was trained on mini-batches and then evaluated on a validation set. Eventually, at each epoch, the training loss, validation loss, training accuracy, and validation accuracy were recorded. To ensure generalization, we coded model checkpointing after every epoch, where the best-performing model based on validation accuracy was saved to be later used during fusion.

The training objective was to minimize the categorical cross-entropy loss, which is well-suited for multi-class classification. The loss function used is the categorical cross-entropy loss, which is defined as:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{6}$$

where,

- $C$ is the number of classes (4 in our case),
- $y_i$ is the ground truth label (one-hot encoded),
- $\hat{y}_i$ is the predicted softmax probability for class $i$.

Also, we employed the Adam optimizer with a learning rate of 1e-4 and no additional momentum terms[31]. The rest of the training summary is provided in Table 5.

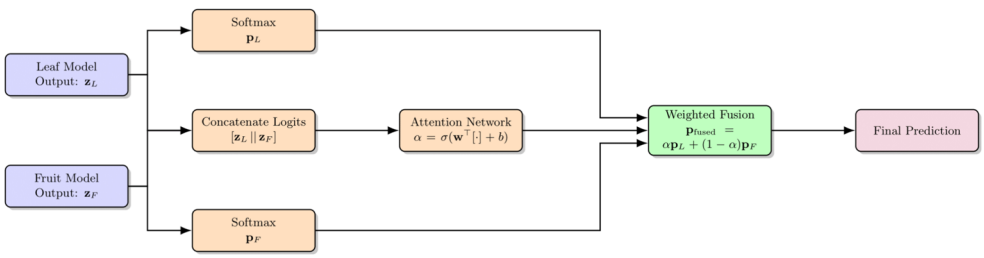



**Fig. 5**. Workflow of the proposed modality attention fusion mechanism.

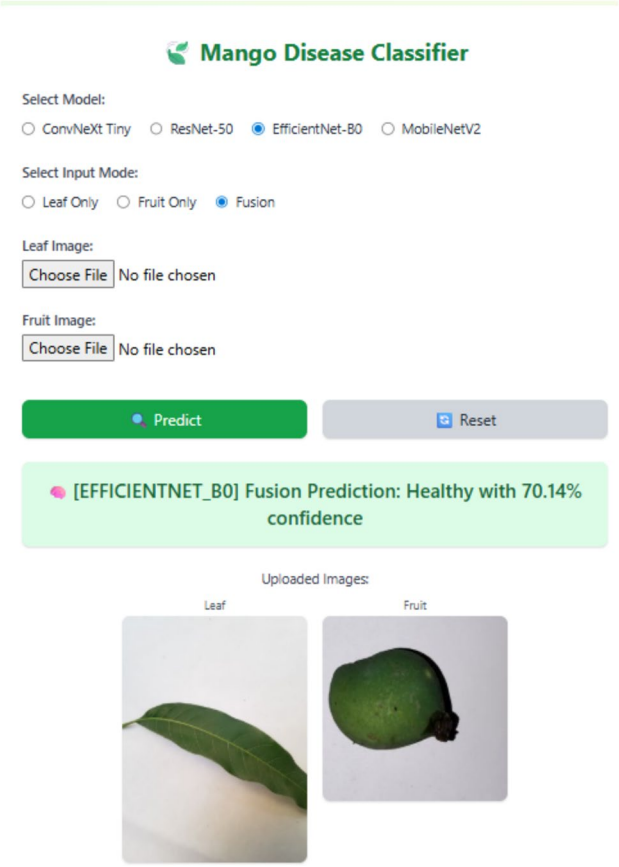



**Fig. 6**. Snapshot of the user interface for the Django-based prediction App.

| Parameter | Value |
|---|---|
| Batch size | 32 |
| Epochs | 10 |
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-4}$ |
| Loss function | CrossEntropyLoss |
| Device | GPU |
| Model saving strategy | Best model by validation accuracy |

**Table 5**. Training configuration and hyperparameters.

## System configuration

Image-based classification requires a high-tech system, and for that, we utilized the cloud-based free resource available in the form of Google Colab, which provides GPU access for a few hours daily to accelerate computation. The detailed specifications documented during the training process are mentioned in Table 6

## Evaluation parameters

To thoroughly evaluate the models in the context of real-world mango disease classification, we selected multiple standard classification metrics whose significance is particularly important when deployed in agriculture.

*Accuracy*

Accuracy is defined as the measure of the proportion of correctly predicted samples out of the total number of predictions. In other words, it reflects the system's ability to correctly identify a wide range of diseases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

| Component | Specification |
|---|---|
| Operating system | Ubuntu 22.04.4 LTS |
| CPU model | Intel(R) Xeon(R) CPU @ 2.00GHz |
| Total CPU cores | 2 |
| Total RAM | 12.67 GB |
| Used RAM | 3.60 GB |
| Total disk storage | 112.64 GB |
| Used disk storage | 43.14 GB |
| Free disk space | 69.49 GB |
| GPU model | NVIDIA Tesla T4 |
| Total GPU memory | 14.74 GB |
| Used GPU memory | 4.45 GB |

**Table 6**. System configuration.

*Precision*
Precision is the proportion of the predicted positive cases that are actually correct. It means, out of all the cases that are predicted as a specific class, how many were actually correct, and it is critical when false positives are costly. For example, if the system incorrectly classifies a healthy fruit as Anthracnose, a farmer might discard it or apply unnecessary chemical treatments.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

*Recall*
It is the proportion of actual positive cases that were correctly identified, which means that out of all the actual cases of a disease, how many were correctly identified. Recall is vital when missing or misclassifying a disease is risky, especially in the case of a contagious disease like Bacterial Black Spot, which can rapidly spread across trees if left untreated.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

*F1 score*
F1 Score is the harmonic mean of precision and recall, where it balances both false positives and false negatives. It provides a balanced view by avoiding both Type I errors (treating healthy fruit unnecessarily) and Type II errors (missing actual infections).

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

*Cohen's Kappa*
Cohen's Kappa is a measure of agreement between the predicted and actual labels, where a high Kappa suggests that the model isn't just guessing or exploiting an imbalanced class distribution.

$$p_o = \frac{\sum_{i=1}^{k} n_{ii}}{N}, \quad p_e = \sum_{i=1}^{k} \left( \frac{n_{i+} \cdot n_{+i}}{N^2} \right) \tag{11}$$

where,

- $k$ is the number of classes,
- $n_{ii}$ is the number of samples correctly predicted for class $i$,
- $n_{i+}$ is the total actual number of samples in class $i$ (i.e., the sum of row $i$),
- $n_{+i}$ is the total predicted number of samples in class $i$ (i.e., the sum of column $i$),
- $N$ is the total number of samples.

*Matthews correlation coefficient*
Matthews Correlation Coefficient (MCC) is known for handling imbalanced classes well, where it provides a single summary metric that balances true/false positives and negatives, which offers a more informative measure than accuracy in many cases.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

*ROC-AUC*
The ROC-AUC (Receiver Operating Characteristic - Area Under Curve) score evaluates the model's ability to separate the disease classes correctly. It reflects how well the model distinguishes between classes across all decision thresholds.

$$\text{AUC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \left(1 - \frac{FP}{FP+TN}\right)\right) \tag{13}$$

*Balanced error rate*
BER is a measure of the average misclassification rate across all classes. It helps mitigate bias toward majority classes and provides a clearer picture of how the model performs across both common and rare diseases.

$$\text{BER} = \frac{1}{2}\left(\frac{FP}{FP+TN} + \frac{FN}{FN+TP}\right) \tag{14}$$

### K-fold cross-validation
These typical evaluations were not enough to test the generalization capability and consistency of the models across different data splits, so we employed K-Fold cross-validation, where the dataset was stratified to maintain class distribution in each fold. We tested the individual models using 10 folds and the fusion models using 5 folds. During individual model training, the models were trained on ten different train-validation splits, and five during fusion models. Later, the average performance was recorded, as shown in the Section 4. This approach ensures that the results are not biased by a single train-test split and provides a more robust estimate of model performance.

## Result and discussion
This section presents a comprehensive and comparative analysis of the proposed mango disease classification framework, encompassing both individual pre-trained CNN models and fusion-based strategies. The performance of each model during training was evaluated on a validation set and later on a separate test set using diverse metrics. The results are organized into two major components: the first is the performance of individual leaf and fruit models, and the second is fusion through dynamic modality attention.

### Individual model performance
Since we are working with two separate datasets, leaf and fruit, the models were trained and evaluated independently on each. Accordingly, their performance will also be discussed in separate sections. For the results discussed in the subsequent sections, Tables 7, 9, and 12 report the performance of the best epoch (highest validation accuracy) from a single training run (out of 10 epochs).

*Leaf-based classification*
The first phase of evaluation focused on individual CNN architectures trained exclusively on mango leaf images. These models were assessed on a held-out test set using a comprehensive suite of performance metrics to avoid any possible bias. A summary of the results is provided in Table 7, while the respective confusion matrices are collectively visualized in Fig. 7, and according to it, all four CNN architectures demonstrated high classification performance on the leaf dataset. Among them, MobileNetV2 achieved the highest accuracy (98.88%, along with the strongest F1 Score (0.9863) and lowest Balanced Error Rate (0.0119), indicating a well-calibrated balance between sensitivity and specificity across all disease classes. This is especially notable given MobileNetV2's lightweight architecture, suggesting its suitability for real-time deployment on edge devices such as smartphones or UAV-based monitoring systems.

EfficientNet-B0 and ConvNeXt also produced strong performances, with F1 scores of 0.9778 and 0.9738, respectively. EfficientNet-B0's combination of relatively high accuracy (98.32%) and low BER (0.0214) underscores its robustness, particularly for subtle disease classes with limited visual features. ConvNeXt exhibited the highest ROC-AUC (0.9995), confirming its exceptional capacity to distinguish between classes across all thresholds, though its F1 and MCC were slightly lower than those of MobileNetV2 and EfficientNet-B0. But while evaluating the 10-fold cross-validation on the Leaf dataset, as shown in Table 8, the ConvNext model performed exceptionally well in almost every metric.

Meanwhile, ResNet-50, which was widely used in prior literature, showed comparatively lower performance with an F1 score of 0.9689 and the highest BER (0.0339) among the group. Its performance, while still strong,

| Model | Acc. | Prec. | Rec. | F1 | BER | AUC | Kappa | MCC |
|---|---|---|---|---|---|---|---|---|
| ResNet-50 | 0.9776 | 0.9722 | 0.9661 | 0.9689 | 0.0339 | 0.9993 | 0.9616 | 0.9617 |
| MobileNetV2 | **0.9888** | **0.9845** | **0.9881** | **0.9863** | **0.0119** | **0.9994** | **0.9809** | **0.9809** |
| EffNet-B0 | 0.9832 | 0.9771 | 0.9786 | 0.9778 | 0.0214 | 0.9986 | 0.9713 | 0.9713 |
| ConvNeXt | 0.9804 | 0.9771 | 0.9706 | 0.9738 | 0.0294 | 0.9995 | 0.9664 | 0.9665 |

**Table 7**. Performance comparison of individual models on the Mango Leaf dataset. Bold values show the best performance.

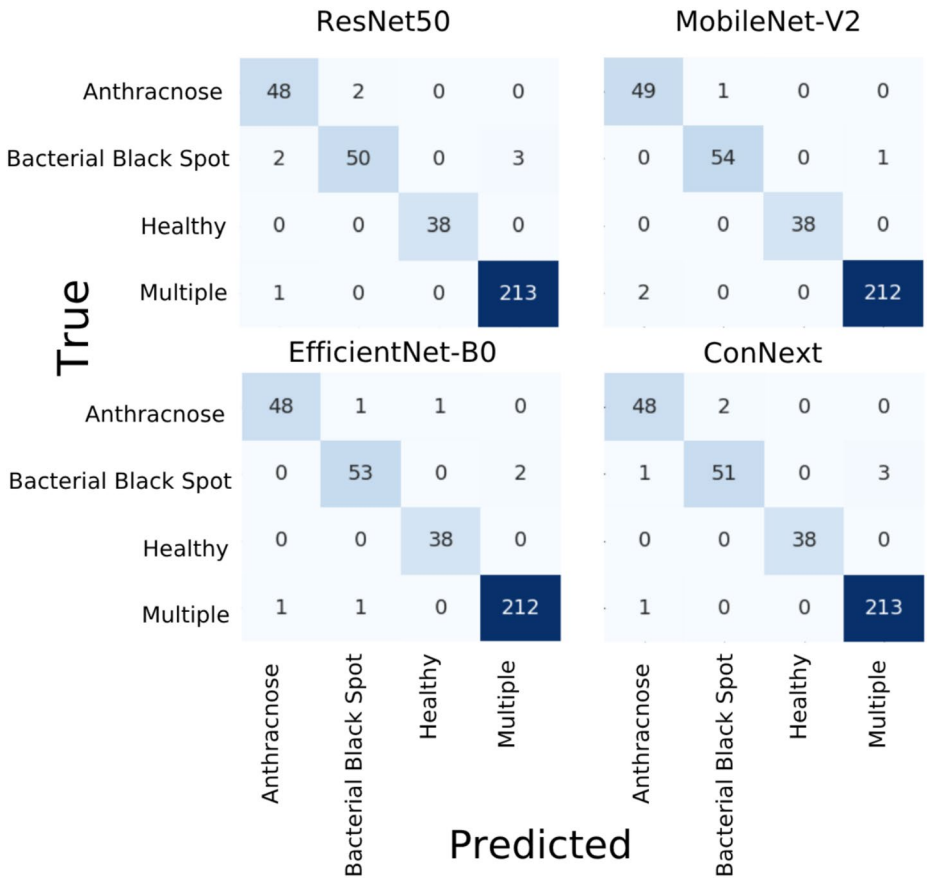



**Fig. 7**. Confusion matrix for the leaf dataset.

| Model | Accuracy | Precision | Recall | F1 | BER | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| ResNet-50 | 0.9860 ± 0.0083 | 0.9825 ± 0.0096 | 0.9806 ± 0.0145 | 0.9811 ± 0.0114 | 0.0194 ± 0.0145 | 0.9762 ± 0.0135 | 0.9763 ± 0.0134 |
| MobileNetV2 | 0.9842 ± 0.0103 | 0.9804 ± 0.0145 | 0.9780 ± 0.0145 | 0.9789 ± 0.0138 | 0.0220 ± 0.0145 | 0.9735 ± 0.0165 | 0.9736 ± 0.0164 |
| EffNet-B0 | 0.9800 ± 0.0083 | 0.9700 ± 0.0133 | 0.9752 ± 0.0118 | 0.9721 ± 0.0114 | 0.0248 ± 0.0118 | 0.9662 ± 0.0130 | 0.9664 ± 0.0128 |
| ConvNeXt-Tiny | **0.9890 ± 0.0075** | **0.9854 ± 0.0103** | **0.9839 ± 0.0108** | **0.9844 ± 0.0102** | **0.0161 ± 0.0108** | **0.9813 ± 0.0125** | **0.9814 ± 0.0124** |

**Table 8**. 10-fold cross-validation results (Mean ± Standard deviation) using the Leaf dataset. Bold values show the best performance.

| Model | Acc. | Prec. | Rec. | F1 | BER | AUC | Kappa | MCC |
|---|---|---|---|---|---|---|---|---|
| ResNet-50 | 0.9824 | 0.9778 | 0.9738 | 0.9758 | 0.0262 | 0.9925 | 0.9750 | 0.9750 |
| MobileNetV2 | **0.9965** | **0.9974** | **0.9931** | **0.9952** | **0.0069** | **0.9980** | **0.9950** | **0.9950** |
| EffNet-B0 | 0.9930 | 0.9904 | 0.9904 | 0.9904 | 0.0096 | 0.9922 | 0.9900 | 0.9900 |
| ConvNeXt | 0.9824 | 0.9824 | 0.9745 | 0.9781 | 0.0255 | 0.9967 | 0.9750 | 0.9752 |

**Table 9**. Performance of individual models on the Mango Fruit dataset. Bold values show the best performance.

suggests that more modern or specialized architectures may provide superior generalization on high-resolution, fine-grained mango leaf datasets.

*Fruit-based classification*

Following the evaluation of leaf-based models, we assessed the performance of CNN architectures trained solely on the mango fruit image dataset. The models were evaluated using the same performance metrics, and the summarized results are reported in Table 9, with confusion matrices illustrated in Fig. 8.

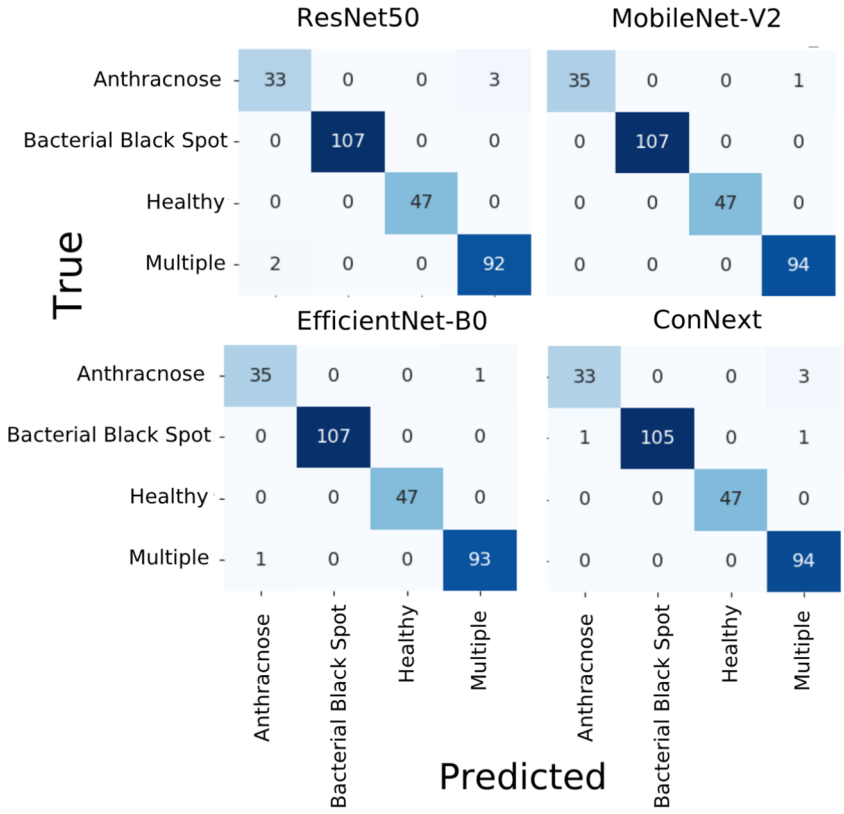



**Fig. 8**. Confusion matrix for the fruit dataset.

| Model | Accuracy | Precision | Recall | F1-Score | BER | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| ResNet-50 | 0.9856 ± 0.0088 | 0.9815 ± 0.0122 | 0.9786 ± 0.0149 | 0.9794 ± 0.0123 | 0.0214 ± 0.0149 | 0.9793 ± 0.0125 | 0.9795 ± 0.0124 |
| MobileNetV2 | 0.9861 ± 0.0068 | 0.9814 ± 0.0110 | 0.9792 ± 0.0130 | 0.9797 ± 0.0104 | 0.0208 ± 0.0130 | 0.9801 ± 0.0097 | 0.9802 ± 0.0096 |
| EffNet-B0 | **0.9872 ± 0.0084** | 0.9809 ± 0.0120 | **0.9842 ± 0.0117** | **0.9821 ± 0.0102** | **0.0158 ± 0.0117** | **0.9817 ± 0.0120** | **0.9818 ± 0.0119** |
| ConvNeXt-Tiny | 0.9861 ± 0.0058 | **0.9823 ± 0.0098** | 0.9788 ± 0.0130 | 0.9798 ± 0.0087 | 0.0212 ± 0.0130 | 0.9800 ± 0.0082 | 0.9802 ± 0.0081 |

**Table 10**. 10-Fold cross-validation performance (Mean ± Standard deviation) of models for the Fruit dataset. Bold values show the best performance.

Here, once again, the MobileNetV2 emerged as the top-performing architecture. It achieved an outstanding accuracy of 99.65%, F1 score of 0.9952, and the lowest Balanced Error Rate (0.0069) across all fruit-based models, which is also supported by the 10-Fold cross validation as shown in Table 10. These results indicate that MobileNetV2 not only maintained its superior classification capacity from the leaf domain but also generalized well to fruit images, reinforcing its potential for lightweight, field-deployable inference.

EfficientNet-B0 also exhibited excellent performance, with a perfectly balanced precision and recall of 0.9904 and a high MCC of 0.9900. This suggests strong class-wise reliability, especially for nuanced diseases with visual manifestations on mango fruit, such as Bacterial Black Spot or Cutting Weevil.

While ResNet-50 and ConvNeXt also performed well, their results were slightly lower compared to MobileNetV2 and EfficientNet-B0. ConvNeXt delivered a strong ROC-AUC score of 0.9967, the highest in the fruit group, indicating excellent ranking capability for distinguishing between disease classes. However, its recall and BER suggest slightly more false negatives, which may stem from subtle fruit disease symptoms or variations in lighting and occlusion in the dataset.

## Dynamic attention based fusion

To further enhance the robustness, generalizability, and practical integration of mango disease classification, we implemented a Modality Attention Fusion (MAF) strategy. This approach learns dynamic, per-sample attention weights for each modality (leaf and fruit) and fuses their logits accordingly. Unlike static fusion methods, this mechanism adapts based on the confidence and discriminative power of each modality, ensuring context-aware decision-making. Thus, a total of 218 pairs were created, and Table 11 lists their distribution among the classes.

The comparative results of modality attention fusion across different base architectures are presented in Table 12, and Fig. 9 shows an attention dynamics visualization that highlights how the learned weights vary across

| Class | Number of pairs |
|---|---|
| Anthracnose | 34 |
| Bacterial black spot | 54 |
| Healthy | 37 |
| Multiple | 93 |
| Total | 218 |

**Table 11**. Distribution of paired samples across classes.

| Models | Acc. | Prec. | Rec. | F1 | BER | AUC | Kappa | MCC |
|---|---|---|---|---|---|---|---|---|
| EffNet-B0 | **0.9908** | **0.9902** | **0.9907** | **0.9903** | **0.0093** | **0.9996** | **0.9870** | **0.9870** |
| MobileNetV2 | 0.9771 | 0.9710 | 0.9769 | 0.9726 | 0.0231 | 0.9993 | 0.9674 | 0.9680 |
| ResNet-50 | 0.9679 | 0.9573 | 0.9676 | 0.9593 | 0.0324 | 0.9980 | 0.9545 | 0.9559 |
| ConvNeXt | 0.8670 | 0.8778 | 0.8945 | 0.8699 | 0.1055 | 0.9873 | 0.8172 | 0.8288 |

**Table 12**. Performance of the fusion models across different backbones. Bold values show the best performance.

classes and samples, where it yielded highly promising results, especially when combined with EfficientNet-B0, which achieved the highest accuracy (99.08%), a perfect ROC-AUC (0.9996), and a minimal BER (0.0093). These results clearly demonstrate the synergy between modalities when guided by an attention-based mechanism that learns how to weigh features contextually. Particularly, in Fig. 9, you can see that the X-axis represents the sample index and the Y-axis indicates the learned attention weight (importance given to either leaf (blue line) or fruit (red line) features).

Since the paired samples were a total of 218 and it was less practical to visualize in full, we limited the visualization to 50 representive samples, where the ResNet50 (top left) and ConvNeXt (bottom right) demonstrated adaptive behavior by initially assigning higher weights to leaf features but later shifting focus to fruit features for certain samples, which indicates their responsiveness to disease-specific visual cues. EfficientNet-B0 (top right) maintains a balanced attention distribution between leaf and fruit modalities; however, in contrast, MobileNetV2 (bottom left) shows a strong and stable preference for leaf features throughout, implying its limited modality flexibility.

The MobileNetV2 fusion model also exhibited strong performance with an accuracy of 97.71% and a balanced error rate of 0.0231, reaffirming its consistency across both individual and fused configurations. The ResNet-50 fusion configuration, while solid, showed slightly reduced performance with a BER of 0.0324, suggesting that although the model benefited from modality fusion, it may not have leveraged modality attention as effectively as other architectures.

In contrast, the ConvNeXt-based fusion model underperformed relative to its individual configuration, achieving the lowest F1 score (0.8699) and the highest BER (0.1055). This drop indicates that while ConvNeXt is capable in unimodal tasks, its fusion performance may be limited by incompatibility with the dynamic weighting mechanism, possibly due to its architecture being tuned for hierarchical visual patterns that do not align well across modalities.

Table 13 illustrates results using 5-fold validation of the fusion models. Results show the better performance of fusion models with improved accuracy and smaller standard deviation values across various evaluation metrics.

## Out-of-distribution generalization testing

As we mentioned earlier, we developed an app to test the real-world application of this research. For this, we performed out-of-distribution testing using real-world mango fruit and leaf images collected from sources outside of the original dataset. For OOD testing, the labeled images were downloaded from Google with different qualities and backgrounds. These images were diverse and presented different lighting conditions, backgrounds, and disease severity levels, which allowed us to assess the model's generalization to unseen conditions. Specifically for 'Bacterial Black Spots', the images were of extremely poor quality, both during training and then on testing, which is why it did not perform well in OOD. For the leaf images, we relied on the publicly available Mendeley dataset[32], and for the fruit images, we used various images available through the Google Search engine. We also employed different manual augmentations and tested around 10 images per class on each model, making a total of 240 images that we tested manually in three modes (only leaf, only fruit, and fusion). The results are presented in the Table 14.

## State-of-the-art comparison

In order to validate the effectiveness of the proposed models, Table 15 summarizes several previous studies that have used custom-merged or multi-source datasets to improve generalizability and performance. Although the ViX-MangoEFormer by[33] achieved an accuracy of 99.78% after being trained on a mixed dataset that includes mango leaves along with other crops like tomatoes, this does not undermine the robustness and adaptability of the fusion model.

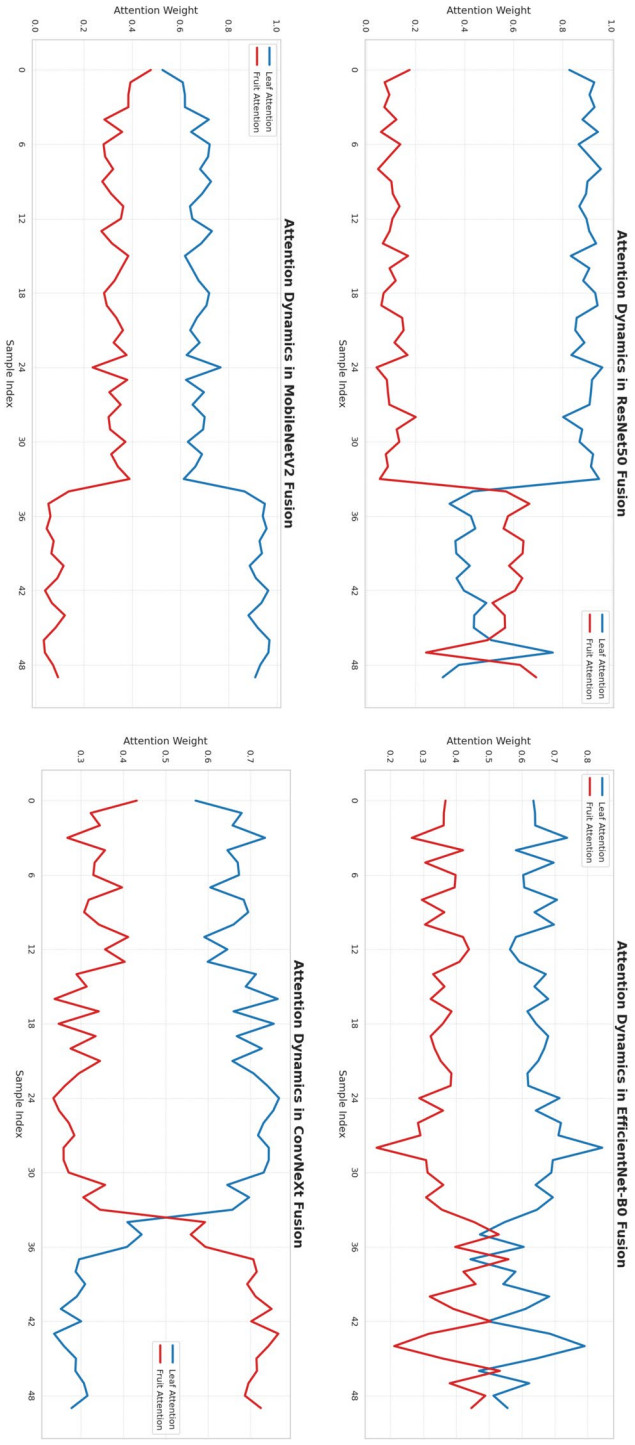



**Fig. 9**. Attention dynamics visualization graphs of different models.

| Model | Accuracy | Precision | Recall | F1-Score | BER | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| EffNet-B0 | 0.9922 ± 0.0074 | 0.9906 ± 0.0103 | 0.9920 ± 0.0076 | 0.9910 ± 0.0093 | 0.0080 ± 0.0076 | 0.9889 ± 0.0105 | 0.9891 ± 0.0103 |
| ResNet-50 | 0.9932 ± 0.0056 | 0.9930 ± 0.0084 | 0.9911 ± 0.0085 | 0.9918 ± 0.0086 | 0.0089 ± 0.0085 | 0.9903 ± 0.0079 | 0.9904 ± 0.0078 |
| MobileNetV2 | **0.9990 ± 0.0022** | **0.9986 ± 0.0031** | **0.9990 ± 0.0022** | **0.9988 ± 0.0027** | **0.0010 ± 0.0022** | **0.9986 ± 0.0031** | **0.9986 ± 0.0031** |
| ConvNeXt-Tiny | 0.9980 ± 0.0027 | 0.9971 ± 0.0039 | 0.9974 ± 0.0036 | 0.9973 ± 0.0038 | 0.0026 ± 0.0036 | 0.9972 ± 0.0038 | 0.9972 ± 0.0038 |

**Table 13**. 5-fold cross-validation performance (Mean ± Standard deviation) of fusion models. Bold values show the best performance.

| Model | Anthracnose | Healthy | Multiple | BBS |
|---|---|---|---|---|
| Leaf dataset | | | | |
| ConvNeXt Tiny | 10 | 9 | 10 | 3 |
| ResNet-50 | 8 | 8 | 10 | 0 |
| EfficientNet-B0 | 3 | 9 | 10 | 10 |
| MobileNetV2 | 1 | 2 | 7 | 2 |
| Fruit dataset | | | | |
| ConvNeXt Tiny | 10 | 10 | 10 | 3 |
| ResNet-50 | 9 | 10 | 9 | 8 |
| EfficientNet-B0 | 10 | 10 | 10 | 4 |
| MobileNetV2 | 10 | 10 | 10 | 7 |
| Fusion results | | | | |
| ConvNeXt Tiny | 10 | 9 | 10 | 4 |
| ResNet-50 | 10 | 10 | 10 | 0 |
| EfficientNet-B0 | 10 | 9 | 10 | 10 |
| MobileNetV2 | 10 | 8 | 10 | 2 |

**Table 14**. Manual testing accuracy per class (10 images each) on Leaf, Fruit, and Fusion inputs.

| References | Method / Model | Dataset(s) Used | Accuracy (%) |
|---|---|---|---|
| [33] | ViX-MangoEFormer | MLD24 (25,530 images) | 99.78 |
| [34] | Enhanced Xception TL model | MLD24 (subset of 5491 images) | 98.00 |
| [35] | ResNet-50 (Severity Estimation) | SenMangoFruitDDS | 97.82 |
| [36] | ConvNeXt + ViT | MangoLeafBD + SenMangoFruitDDS | 98.40 |
| Proposed | MobileNetV2 (Leaf Only) | MLD24 (custom merged) | 98.88 |
| Proposed | MobileNetV2 (Fruit Only) | SenMangoFruitDDS (custom merged) | 99.65 |
| Proposed | EffNet-B0 (Fusion) | Leaf + Fruit + other sets (merged) | 99.08 |

**Table 15**. Performance comparison with existing approaches.

| Augmentation | ResNet-50 | MobileNetV2 | EfficientNet-B0 | ConvNeXt |
|---|---|---|---|---|
| Leaf dataset | | | | |
| Class-aware Aug. | 0.9776 | 0.9888 | 0.9832 | 0.9804 |
| No Augmentation | 0.9832 | 0.9805 | 0.9832 | 0.9860 |
| Common Aug. | 0.9860 | 0.9860 | 0.9888 | 0.9916 |
| Fruit dataset | | | | |
| Class-aware Aug. | 0.9824 | 0.9965 | 0.9930 | 0.9824 |
| No Augmentation | 0.9859 | 0.9894 | 0.9965 | 0.9824 |
| Common Aug. | 0.9754 | 0.9824 | 0.9894 | 0.9930 |

**Table 16**. Ablation study of different augmentation strategies on individual leaf and fruit datasets (accuracy).

## Ablation study

In order to assess the role of augmentation, we compared three strategies :

1. No augmentation at all,
2. A single augmentation pipeline applied uniformly to all classes, and
3. Class-aware augmentation with class-specific pipelines.

These ablation studies were evaluated only across eight individual models (combining both leaf and fruit datasets) because the fusion models operate on the outputs of pre-trained individual models. Results are summarized in Table 16, where we witnessed minimal difference among them, likely due to the dataset diversity and the robustness of pre-trained backbones.

### Practical implications of the proposed framework

i. *Early disease detection and intervention:* The frameworks capable of early disease detection can help farmers take timely actions to avoid disease spread and reduce yield loss
ii. *Timely decision support:* Currently, the proposed approach offers web-based disease detection. It can be integrated into a mobile app and provide farmers with real-time disease alerts through a smartphone. It can make the framework accessible even to smallholders, which increases farmers' decision-making capabilities.
iii. *Economic benefits:* By reducing losses from diseases, the framework ensures higher market value, protects farmer income, and enhances supply chain stability.
iv. *Scalability for smart agriculture:* The framework can be expanded across other types of diseases, as well as other plants. In addition, IoT-based sensors and drones can be integrated for large-scale monitoring, enabling automation in plantation management.
v. *Food Security & export competitiveness:* Healthier mango crops improve food security locally and ensure compliance with international phytosanitary standards, strengthening the export potential of mango-producing countries.
vi. *Data-driven insights:* Aggregated disease data over time can guide breeding programs, policymaking, and agricultural extension services, promoting long-term sustainable cultivation practices.

### Limitations and future work

While the proposed multimodal framework achieved impressive performance in mango disease classification, certain limitations should be acknowledged. First, the datasets used, although supplemented with out-of-distribution samples, lacked broad geographic and seasonal diversity, potentially limiting the model's generalizability across different cultivation regions and rare disease cases. Additionally, environmental factors such as extreme lighting, background clutter, or partially visible symptoms may still impact prediction accuracy. Another practical limitation is the reliance on paired fruit and leaf images for fusion, which may not always be feasible in real-world scenarios where only one modality is available. In some instances, the user may submit images of two different classes, which the models can't predict because the dataset assumes class-consistent leaf and fruit samples. In practice, mismatches (e.g., healthy leaf but diseased fruit) may occur, and such cases were not modeled in training and could reduce applicability in real-world settings. Also, in the dataset, the BBS (Bacterial Black Spot) images fed during training were of extremely low quality, which limited their performance during OOD testing. As a result, the model did not perform well in generalized testing. Adding higher-quality images in the future could help mitigate this issue. Furthermore, the attention-based fusion models are relatively resource-intensive, making deployment on low-power or mobile devices more challenging.

Looking ahead, future work can explore more flexible, cost-efficient, and scalable solutions, such as incorporating temporal or multisensor data, such as thermal or hyperspectral imagery, which could further enhance disease detection accuracy. Developing methods for unpaired or single-modality fusion would improve applicability in field conditions with incomplete data. Active learning strategies could also be employed to reduce the annotation burden while continually improving model robustness. Moreover, optimizing the model for edge devices can facilitate real-time inference and practical deployment in agricultural settings. Finally, integrating interpretability tools, such as attention visualizations or class activation maps, can provide transparency and foster trust among end-users, including farmers and agronomists.

## Conclusion

This research presented a multimodal deep learning framework for the automated classification of mango diseases using both leaf and fruit images. By training individual CNN models like ResNet-50, MobileNetV2, EfficientNet-B0, and ConvNeXt and integrating their predictions through a Modality Attention Fusion (MAF) strategy, we demonstrated significant improvements in classification performance, generalization, and interpretability. The fusion models dynamically prioritized the more informative modality per sample, enabling robust prediction even in the presence of occlusions, viewpoint variations, or partial symptoms. Our best-performing fusion model was EfficientNet-B0, which achieved a test accuracy of 99.08% and an F1 score of 0.9903, outperforming both unimodal and static fusion baselines. These findings underline the importance of combining visual cues from multiple plant organs and adapting model attention to contextual features for real-world agricultural diagnostics. We also developed a web application using the Django framework to perform out-of-distribution testing, where the models demonstrated remarkable results.

## Data availability

The dataset can be requested from Muhammad Shadab Alam Hashmi (shadab.alam@kfueit.edu.pk).

Received: 30 July 2025; Accepted: 27 October 2025
Published online: 26 November 2025

## References

1. Lebaka, V. R., Wee, Y.-J., Ye, W. & Korivi, M. Nutritional composition and bioactive compounds in three different parts of mango fruit. *Int. J. Environ. Res. Publ. Health* **18**(2), 741. https://doi.org/10.3390/ijerph18020741 (2021).
2. Hossain, M. A., Sakib, S., Abdullah, H. M. & Arman, S. E. Deep learning for mango leaf disease identification: A vision transformer perspective. *Heliyon* **10**(8), 36361. https://doi.org/10.1016/j.heliyon.2024.e36361 (2024).
3. Deliana, Y., Fatimah, S. & Charina, A. Marketing and value chain of 'gedong gincu' mango with its labeling and packaging. in *Acta Horticulturae*, vol. 1183, 319–326 (2017). https://doi.org/10.17660/ActaHortic.2017.1183.53

4. Maldonado-Celis, M. E. et al. Chemical composition of mango (*Mangifera indica* l.) fruit: Nutritional and phytochemical compounds. *Front. Plant Sci.* **10**, 1073. https://doi.org/10.3389/fpls.2019.01073 (2019).
5. Bana, J. K., Sharma, H., Chavan, S. M., Sharma, D. K. & Patil, S. J. Assessment of losses caused by major insect-pests and diseases of mango (*Mangifera indica* l.) under humid tropics. *Pest Manag. Hortic. Ecosyst.* **29**(2), 213–220 (2024).
6. Nazir, R., Abbas, A., Ahmad, M., Akram, W. & Iqbal, M. Occurrence and severity of mango decline disease in Punjab, Pakistan. *Sarh. J. Agric.* **36**(2), 556–562. https://doi.org/10.17582/journal.sja/2020/36.2.556.562 (2020).
7. Karad, S., Sonavane, S., Kshirsagar, S., Dalvi, S. & Ahire, J. Mobile-based mango disease detection using image classification techniques. *Int. Res. J. Modern. Eng. Technol. Sci.***7**(4) https://doi.org/10.56726/IRJMETS73768 (2025).
8. Batool, F., Akhtar, M. N., Azhar, M. & Tariq, M. Ensembled attennet: A novel deep learning approach for mango leaf disease detection. *Phys. Educ. Health Soc. Sci.* **3**(1), 162–171. https://doi.org/10.63163/jpehss.v3i1.135 (2025).
9. Kunduracioglu, I. & Pacal, I. Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases. *J. Plant Dis. Prot.* **131**(3), 1061–1080 (2024).
10. Kunduracioglu, I. Utilizing resnet architectures for identification of tomato diseases. *J. Intell. Decis. Making Inf. Sci.* **1**, 104–119 (2024).
11. Kunduracioglu, I. CNN models approaches for robust classification of apple diseases. *Comput. Decis. Making Int. J.* **1**, 235–251. https://doi.org/10.59543/comdem.v1i.10957 (2024).
12. Pacal, I. et al. A systematic review of deep learning techniques for plant diseases. *Artif. Intell. Rev.* **57**(11), 304 (2024).
13. Lee, C. P., Lim, K. M., Song, Y. X. & Alqahtani, A. Plant-CNN-VIT: Plant classification with ensemble of convolutional neural networks and vision transformer. *Plants* **12**(14), 2642. https://doi.org/10.3390/plants12142642 (2023).
14. Tonmoy, M. R., Hossain, M. M., Dey, N. & Mridha, M. F. Mobileplantvit: A mobile-friendly hybrid vit for plant disease classification. arXiv preprint arXiv:2503.16628https://doi.org/10.48550/arXiv.2503.16628 (2025).
15. Aboelenin, S. et al. A hybrid framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer. *Complex Intell. Syst.* **11**, 142. https://doi.org/10.1007/s40747-024-01764-x (2025).
16. Kaya, Y. & Gürsoy, E. A novel multi-head CNN design to identify plant diseases using the fusion of RGB images. *Eco. Inform.* **75**, 101998. https://doi.org/10.1016/j.ecoinf.2023.101998 (2023).
17. Wang, X. et al. A multimodal data fusion and embedding attention mechanism-based method for eggplant disease detection. *Plants* **14**(5), 786. https://doi.org/10.3390/plants14050786 (2025).
18. Shakib, M. M. H., Mustofa, S. & Ahad, M. T. MLD24: An image dataset for mango leaf disease detection. https://doi.org/10.17632/6dvpywm2m2.1 .
19. Sanahuja, G., Ploetz, R. C., Lopez, P., Konkol, J. L., Palmateer, A. J. & Pruvost, O. Bacterial canker of mango, *Mangifera indica*, caused by *Xanthomonas citri* pv. *mangiferaeindicae*, confirmed for the first time in the americas. Plant Disease **100**(12), 2532 (2016) https://doi.org/10.1094/PDIS-03-16-0412-PDN
20. Faye, D., Diop, I., Mbaye, N., Diedhiou, M. & Dione, D. SenMangoFruitDDS. https://doi.org/10.17632/jvszp9cbpw.4 .
21. Basak, S. K. Mango fruit early diseases detection. Kaggle (2025). https://doi.org/10.34740/KAGGLE/DSV/11848661 . https://www.kaggle.com/dsv/11848661
22. Mohanty, S. P., Hughes, D. P. & Salathé, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **7**, 1419. https://doi.org/10.3389/fpls.2016.01419 (2016).
23. Ferentinos, K. P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **145**, 311–318. https://doi.org/10.1016/j.compag.2018.01.009 (2018).
24. Veling, S. Mango disease detection by using image processing. *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)* (2019) https://doi.org/10.22214/IJRASET.2019.4624
25. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90 .
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520 (2018). https://doi.org/10.1109/CVPR.2018.00474 .
27. Kunduracıoğlu, İ & Paçal, İ. Deep learning-based disease detection in sugarcane leaves: evaluating efficientnet models. *J. Operat. Intell.* **2**(1), 321–235 (2024).
28. Tan, M. & Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. in *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Proceedings of Machine Learning Research, vol. 97, 6105–6114. PMLR, (2019). https://doi.org/10.48550/arXiv.1905.11946 . https://proceedings.mlr.press/v97/tan19a.html
29. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. A convnet for the 2020s. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11966–11976 (2022). https://doi.org/10.1109/CVPR52688.2022.01167
30. Zhou, T., Ruan, S., Guo, Y. & Canu, S. A multi-modality fusion network based on attention mechanism for brain tumor segmentation. in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* 377–380 (2020). https://doi.org/10.1109/ISBI45749.2020.9098392
31. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
32. Ali, S., Ibrahim, M., Ahmed, S. I., Nadim, M., Mizanur, M. R., Shejunti, M. M. & Jabid, T. MangoLeafBD Dataset. https://doi.org/10.17632/hxsnvwty3r.1
33. Noman, A. A. et al. ViX-MangoEFormer: an enhanced vision transformer-efficientformer and stacking ensemble approach for mango leaf disease recognition with explainable artificial intelligence. *Computers* **14**(5), 171. https://doi.org/10.3390/computers14050171 (2025).
34. Sultan, T. et al. LeafDNet: transforming leaf disease diagnosis through deep transfer learning. *Plant Dir.* **9**(2), 70047. https://doi.org/10.1002/pld3.70047 (2025).
35. Faye, D. et al. Mango fruit diseases severity estimation based on image segmentation and deep learning. *Discov. Appl. Sci.* **7**, 143. https://doi.org/10.1007/s42452-025-06550-z (2025).
36. Alamri, F. S., Sadad, T., Almasoud, A. S., Aurangzeb, R. A. & Khan, A. Mango disease detection using fused vision transformer with convnext architecture. *Comput. Mater. Contin.* https://doi.org/10.32604/cmc.2025.061890 (2025).


## Acknowledgements

The authors extend their gratitude to the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R746), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

## Author contributions

MM conceived the idea, performed data analysis and wrote the original draft. MSAH performed data curation, formal analysis, and designed methodology. IDN designed methodology, dealt with software and performed visualization. HG acquired the funding for research, and performed visualization and initial investigation. NAS performed investigation, dealt with software and provided resources. IA supervised the study, performed validation and review and edit the manuscript. All authors read and approved the final manuscript.

## Funding
This study is partially funded by the European University of Atlantic and the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R746), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

## Declarations

## Competing interests
The authors declare no competing interests.

## Ethical and consent to participate
Not applicable.

## Consent for publication
Not applicable.

## Additional information
**Correspondence** and requests for materials should be addressed to M.S.A.H. or I.A.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.