



Attention-Fused ConvNeXt and MobileViT Framework with FastSAM Lesion Extraction for Multi-Class Crop Disease Detection

¹Gopal Ghimire, ²Prof. Raj Kumar Thakur, ³Prof. Dr. Gopal Prasad Sharma, ⁴Sagar Sitaula, ⁵Kiran Barakoti, ⁶Drona Prasad Acharya, ⁷Sumina Thapa, ⁸Keshav Kharal

¹MCA Student, Purbanchal University School of Science & Technology (PUSAT), Biratnagar, Nepal

^{2,3}Professor, Purbanchal University School of Science & Technology (PUSAT), Biratnagar, Nepal

^{4,5}Independent Researchers, Nepal

⁶Independent Researcher, Australia

⁷B.Sc. Agriculture Graduate, Agriculture and Forestry University, Faculty of Agriculture, Nepal

⁸Head Technical Assistant (IT), Degree campus, Biratnagar, Nepal

¹ORCID ID: <https://orcid.org/0009-0005-5044-8588>

⁴ORCID ID: <https://orcid.org/0009-0006-1687-6594>

⁵ORCID ID: <https://orcid.org/0009-0005-1250-5342>

⁶ORCID ID: <https://orcid.org/0009-0005-6701-7223>

⁷ORCID ID: <https://orcid.org/0009-0008-5353-3827>

⁸ORCID ID: <https://orcid.org/0009-0006-3518-8930>

DOI: <https://doi.org/10.51583/IJLTEMAS.2026.150800045>

Received: 29 August 2026; Accepted: 03 September 2026; Published: 08 September 2026

ABSTRACT

The problem of crop disease recognition from leaf images is a difficult multi-class visual classification problem due to disease evidence appearing at various spatial scales and the possibility of image content other than disease symptom including healthy tissue, background information, and visually similar symptoms. This paper presents an attention-fused dual-branch framework that combines Fast Segment Anything (FastSAM) based lesion extraction, ConvNeXt based full-leaf representation learning, and MobileViT based lesion-focused representation learning. The experimental study adopted a controlled subset of 4000 images from the 20k+ Multi-Class Crop Disease Images collection and applied class normalization and removed one small class, leaving 41 classes. There were 2777 training images, 588 validation images, and 635 testing images in the final split. The candidate masks were obtained by FastSAM, and then filtered and ranked with lesion-oriented criteria, then the selected region was cropped and refined by performing Lab-space contrast enhancement, robust chromatic deviation, Otsu thresholding and morphological operations. The original leaf was processed by ConvNeXt and the refined ROI processed by MobileViT. They were projected into a common 256-dimensional feature space and aggregated using a learnable two-branch gate that includes entropy regularization. The Fusion Model got 93.07% accuracy and 85.27% macro F1 on the common test set, while ConvNeXt got 89.76% accuracy and 81.55% macro F1, and MobileViT got 72.76% accuracy and 68.07% macro F1 on the common test set. The Fusion Model also achieved a one-versus-rest macro-average AUC of 0.9978. Grouped confusion matrices indicate that the fusion approach reduces off-diagonal errors at the crop-group level. A prototype of a Streamlit deployment was also created to showcase the following features: upload of images, lesion preprocessing, prediction, reporting of confidence, inspection of branches, and image visualization with Grad-CAM. The results support the use of a combination of global and lesion-specific representations rather than a single visual pathway.

Keywords: Crop disease detection, FastSAM, lesion extraction, ConvNeXt, MobileViT, attention fusion, deep learning

INTRODUCTION

Crop diseases must be promptly diagnosed and treated if they are to be controlled in agriculture. Late diagnosis can lead to the spread of disease within a crop, decrease the quality of the harvested crop and complicate the control. Given that the digital photograph also allows for the representation of visible disease symptoms in a compact form and can be used for machine learning models without requiring an expert to manually inspect each of the samples, image-based diagnosis has become an important research direction. The overarching idea driving this work is to frame the problem in terms of real-world crop disease recognition and introduce a fully developed pipeline that can go from an input leaf image to lesion localization, feature learning, classification, and visual explanation.

In image classification, deep learning has significantly revolutionized the field by enabling the learning of representations directly from examples. The early convolutional methods proved that learning visual features in a hierarchical order was effective, and later large-scale architectures developed to learn discriminative representations from large collections of images [7], [13]. These are very helpful for plant disease recognition, as disease symptoms can occur as changes in color, spots, lesions, texture, shape, vein patterns or spatial distribution across a leaf. The disease can also manifest itself in different ways, depending on the severity, lighting, crop variety and image acquisition circumstances. A classifier based on a single visual cue may therefore not perform equally well across all classes.

Transformer-based vision models added another dimension by modelling the relationships between image regions and by offering a way to represent the image in a wider context [8]. At the same time, lightweight transformer-inspired models like MobileViT have been developed that merge the contextual modeling power of the transformer with the spatial structure of convolutional models, but still adapt to constrained settings [10]. The literature review reveals an increasing number of studies that have explored the integration of convolutional and transformer models for crop disease identification and classification, including [1], [2], [3], [5], [15], [16], [19], and [24]. The general idea is that there may be different but complementary information in the local texture and in the wider context.

While these advances have been made, it is not guaranteed that the model will focus on the most informative disease region with direct classification of the entire image. A leaf image may have healthy parts that are so prevalent in the pixel distribution that they mask other parts of the image, background areas, shadows, or other visual features. Hence, the present work introduces an explicit lesion-focused pathway. FastSAM is used to generate the selected region, which is subsequently refined before being passed to the MobileViT branch [11]. The original leaf is kept for ConvNeXt to prevent loss of the general spatial information that could be crucial to identifying classes with similar lesion appearances.

In this paper, then, it is not the replacement but the complementary representation that is the key idea. While ConvNeXt is tasked with handling information from the entire leaf, MobileViT is given a trimmed-down region of interest with lesions. The two feature vectors are aligned and fused together by a learnable attention gate. Rather than assigning a fixed contribution to each branch, the gate produces two normalized weights for each input. It follows the general principle of attention-based adaptive weighting while remaining consistent with the implemented fusion architecture [25]. To prevent early collapse to one branch, entropy regularization is also applied during training of the fusion.

The two component baselines are used intentionally for the evaluation of the experimental contribution. The ConvNeXt baseline represents the best that can be done with the full-leaf information. The performance of the lesion-focused pathway only is reported in the MobileViT baseline. The Fusion Model evaluates whether combining the two branches improves performance over the individual models. The same test set of 635 images is used to evaluate all three models and the same 41-class mapping is used. This is a common evaluation design which is more meaningful than comparing across unrelated public benchmarks.

Also, the paper focuses on deployment, not just offline metrics. The implementation of the study consists of a streamlit prototype that accepts an image, applies the necessary preprocessing, runs the chosen model, returns predictions and confidence information to the user, and generates grad cam explanation images [12]. This

prototype is meant to demonstrate that the research pipeline can be implemented in an interactive workflow, but is not a field-tested agricultural advisory system.

The principal contributions can be summarized as follows:

- A FastSAM-based lesion extraction and refinement workflow that converts an input leaf image into a lesion-focused region for the second branch.
- A dual-branch representation strategy in which ConvNeXt retains full-leaf context and MobileViT emphasizes localized lesion information.
- An attention-gated feature fusion mechanism that learns input-dependent branch weights in a common representation space.
- Gate entropy regularization and staged fusion training are used to improve branch utilization and training stability.
- A controlled comparison of ConvNeXt, MobileViT, and Fusion on the same 41-class test set, including accuracy, macro precision, macro recall, macro F1, weighted F1, ROC-AUC, and grouped confusion-matrix analysis.
- A functional deployment prototype with FastSAM preprocessing, prediction, confidence reporting, branch-wise inspection, and Grad-CAM visualization.

Related Work

Several approaches, including conventional image processing, convolutional neural networks, transfer learning, vision transformers, and hybrid models, have been studied for plant disease recognition. The literature included in the study, shows a gradual shift from single backbone classifiers to systems that try to integrate complementary representations. Zhu et al. explored crop disease identification based on the combination of multiscale convolution and a vision transformer, demonstrating the importance of integration of different representations that have different receptive-field behaviors [1]. This direction is also quite similar to the present work, except that an extra lesion-extraction stage is added and a full-leaf ConvNeXt branch is employed in addition to a lesion-focused MobileViT branch. Gebremedhin et al. also examined the lightweight CNN–MobileViT models used for indigenous crops [17], Chen et al. researched the SinGAN-type and ResNet34 model for plant leaf disease recognition, and Ashurov et al. explored the depth wise CNN, squeeze-and-excitation, and residual connections for plant disease detection [20].

Lightweight transformer designs are especially important for a research system built to be interactive or mobile. In [2] Tonmoy et al. presented a method, called MobilePlantViT, that is designed for the general problem of plant disease image classification and focuses on the trade-off between recognition capability and computational feasibility. Rodrigo et al. studied MobileViT based classification of rice leaf disease in mobile applications [3]. The above studies justify using MobileViT as a branch of the research presented here rather than replacing a full-image pathway but does not suggest that MobileViT is more suitable. The standalone MobileViT model, in the current experiments, even underperforms ConvNeXt and thus should be considered as an additional branch.

Crop disease image classification continued through the use of a convolutional neural network which was the focus of the work in [4] and is again important in the context of crop imagery. It is a newer advancement of the convolutional design called ConvNeXt. Liu et al. [9] proposed an updated convolutional architecture ConvNeXt, which aims to restore the design ideas that had been linked to the transformer models but still preserve the structure of convolution. The current setup chooses ConvNeXt as the network for providing a strong baseline for the full leaf as well as a global representation path to measure lesion focused learning.

Several works that incorporate ConvNeXt or convolutional networks with transformer representations are also found in the study literature. Ergün explored a fusion model of ConvNeXt and Vision transformer for the classification of mango leaf diseases [5]. Aboelenin et al. introduced the hybrid convolutional neural network and vision transformer architecture to detect plant leaf diseases [16]. Huang et al. presented a work that uses ConvNeXt and transformer representations for generalized apple leaf disease classification, named EConv-ViT [19]. The present study builds on the general methodological direction established by these works. When different feature extractors capture a different aspect of the appearance of the disease, hybridization can be useful.

The concept of vision transformers is rooted in the abstraction of images as series of patches and training their relationships [8]. The transformer representation was also shown to be relevant to agriculture image analysis with the work of Borhani et al. on vision transformer based automated plant disease classification [15]. Salman et al. investigated the generalization problem of plant disease classification in the wild using vision transformers and mixture-of-experts concepts [24]. These studies are relevant when interpreting the current results, particularly because the present dataset does not represent uncontrolled field conditions.

Lightweight Convolutional architectures have also had a bearing on the design of efficient agricultural vision systems. Howard et al. [14] introduced MobileNetV3, and Howard et al. introduced the MobileNets that are efficient convolutional networks for mobile vision [23] before that. The models illustrate the long-term goal of using less computation but still representing what is useful. MobileViT builds on these efficiency-focused design principles by combining convolutional and transformer-style processing [10]. The following deployment component of this work fits into this motivation, but not formally as a latency or mobile hardware benchmark, so no claim is made here.

Another line of research is disease recognition by segmentation assistance that is relevant to the proposed method. Within the family of Segmentation Anything, FastSAM has been developed to be faster and can generate candidate segmentation masks for an image [11]. For the current implementation, FastSAM is not considered as a disease-segmentation oracle. Rather, it is used as a candidate-region generator, and then filtered, selected, cropped and processed again through a lesion refinement procedure. Importantly, the method does not assume that FastSAM masks are identical to biologically annotated lesions.

Explainability is also relevant in the case of image classifiers that are designed for practical diagnosis. Grad-CAM generates so-called gradient based visual explanations that identify areas of the image that are important for the model's decision [12]. In this prototype for deployment, Grad-CAM is used to make the predictions visually inspectable. The study presents a branch-wise explanation for the dual architecture, particularly useful due to the fact of having two different views of the same leaf. The explanation does not prove a biological cause, but is meant as a visual diagnostic instrument for checking the action of the learned model.

The data consists of the public Multi-Class Crop Disease Images [26] with 20k+ images. The study also mentions the work on the open image repository related to PlantVillage done by Hughes and Salathé [6]. These resources show the significance of public plant disease image collections. But the results of different sets of data should not be compared simply on the basis of accuracy; the definitions of the classes, the number of images, the type of disease, the sampling procedure, and the environment differ. Current paper therefore lays stress on the internal comparison in controlled way among the three models of the present paper.

Dataset And Experimental Design

A. Source Dataset and Class Preparation

The public 20k+ Multi-Class Crop Disease Images collection [26] is used as the source set for this study. Multiple crop and disease category and naming variations are present in the original material. Class names were normalized before training the model to minimize differences due to capitalization, underscores, spaces and naming conventions. After normalization, 42 common classes were found, the study states. A single class,

labeled “bollrot on Cotton” had four combined images and was dropped due to the small number of images deemed as inadequate for training and evaluation. Hence, an experimental setup with 41 classes is obtained.

The disease and healthy classes for cotton, rice, maize, wheat, sugarcane, and general/pest classes are included in the final class list. The following list is the full list retained in the study American Bollworm on Cotton, Anthracnose on Cotton, Army worm, Bacterial Blight in Rice, Brown spot, Common Rust, Cotton Aphid, Flag Smut, Gray Leaf Spot, Healthy Maize, Healthy Wheat, Healthy cotton, Leaf Curl, Leaf smut, Mosaic sugarcane, Red Rot sugarcane, Red Rust sugarcane, Rice Blast, Sugarcane Healthy, Tungro, Wheat Brown leaf Rust, Wheat Stem fly, Wheat aphid, Wheat black rust, Wheat leaf blight, Wheat mite, Wheat powdery mildew, Wheat scab, Wheat -Yellow- Rust, Wilt, Yellow Rust Sugarcane, bacterial blight in Cotton, bollworm on Cotton, cotton mealy bug, cotton whitefly, maize ear rot, maize fall armyworm, maize stem borer, pink bollworm in cotton, red cotton bug, and thrips on cotton.

The entire raw collection is not directly used in the study. The study, however, developed a controlled 4000 image subset with a maximum of 200 images per class before final allocation. The rationale for this was to maintain a manageable size for the experiment and to minimise differences in class availability. Only the resulting split is reported in the paper and not a fully balanced data set because the actual split is not mathematically equal across all classes.

Table 1: Composition of the experimental dataset

Subset	Images	Role
Training	2777	Model optimization
Validation	588	Model selection and monitoring
Testing	635	Final held-out evaluation
Total	4000	Capped experimental subset
Classes	41	Final classification labels

B. Data Splitting and Paired Inputs

In the study implementation a fixed stratified splitting procedure is used for a random seed of 42. The training ratio is reported as 0.70 and validation ratio as 0.15, while the rest of the samples are assigned for testing. Given that the final number of images is 4000, the numbers of training, validation and test images become 2777, 588, and 635 respectively. The test set is the same for all three models evaluated. This is a common test set which is essential for the experimental design as it ensures that performance differences among the models arise under the same held-out examples rather than from different test populations.

Each original image of leaf is matched with its lesion image in the common name by a common key named filename. This is required because the two branches represent the same instance of leaf. The complete image of each sample should be aligned with the corresponding lesion crop from the same sample. Otherwise, the fusion model could learn an artificial relationship unrelated to the disease label. The paired preparation thus ensures identity of the sample along the two visual pathways.

C. Data Augmentation and Normalization

Standard image preprocessing is used for the underlying implementation, suitable for pretrained vision models. Original images are resized to 224×224 pixels, then normalized based on the ImageNet statistics and sent to the ConvNeXt branch. The MobileViT branch receives cropped images resized to 288×288 pixels, focusing on the lesion regions. The spatial resolution difference is due to the use of branch-specific preprocessing in the study. The full-leaf branch is used to retain the broader structure and the lesion branch uses a higher-resolution input to preserve localized patterns after cropping.

The MobileViT branch also leverages balanced sampling and label smoothing of 0.1. The purpose of balanced sampling is to avoid having larger training support classes dominating the optimization process. Label smoothing helps to prevent the classifier from being too certain about the training labels. Weight decay and gradient clipping are used as regularization and optimization controls. The choices are especially important because the branch focusing only on the lesion lacks the context of the entire leaf, which can make it more sensitive to quality and variation of the extracted region.

PROPOSED METHODOLOGY

A. FastSAM-Based Lesion Candidate Extraction

The lesion-processing stage starts with FastSAM X, an implementation based on the Fast Segment Anything approach [11]. FastSAM is fed the RGB leaf image, and outputs candidate masks. The pipeline handles situations in which the first inference configuration does not return useful masks. The implementation therefore uses different confidence and intersection-over-union parameters for retry attempts and can increase the input size to 1280 pixels. This retry logic is necessary because segmentation failure would otherwise interrupt the downstream lesion-processing branch.

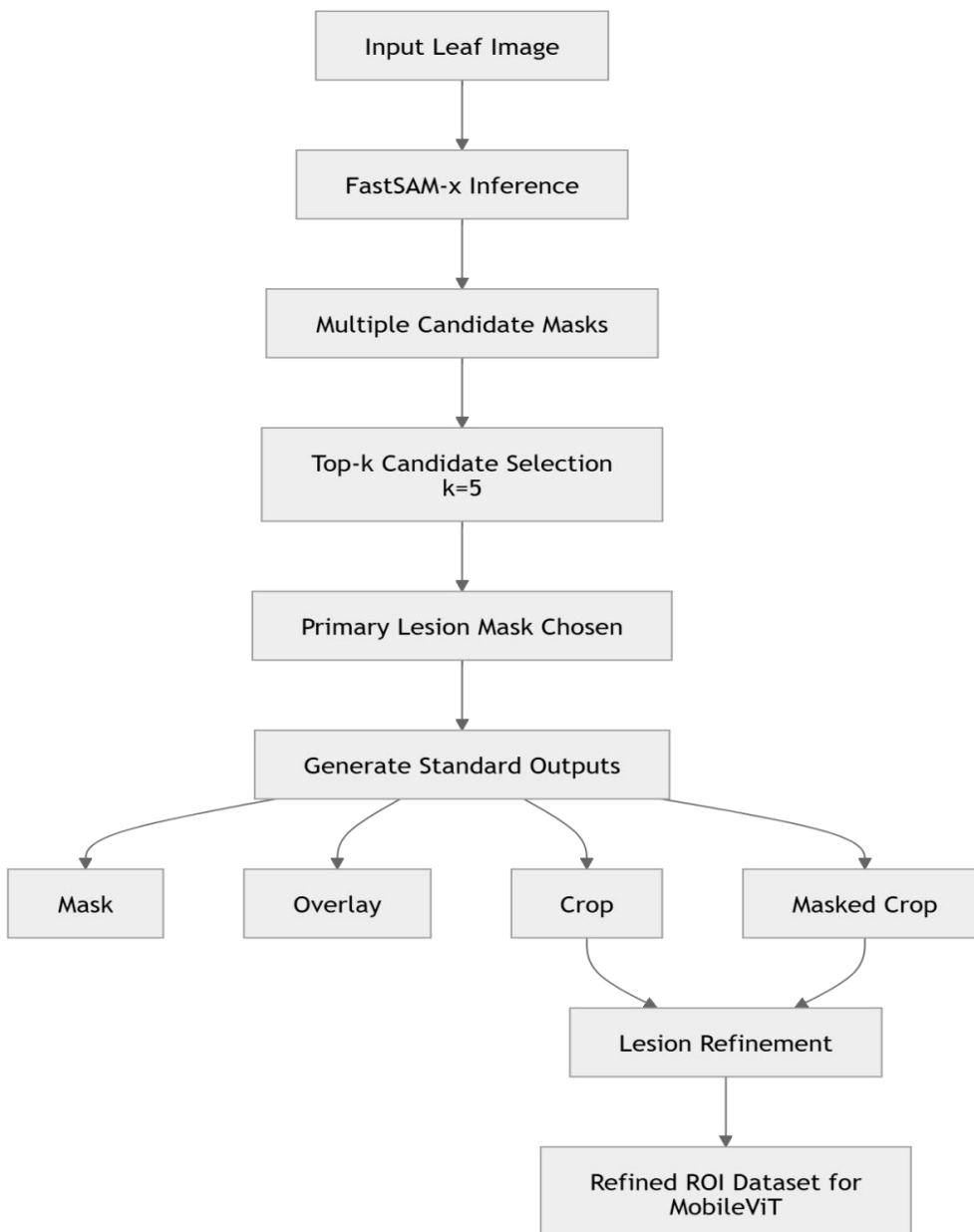


Figure 1: FastSAM-based lesion extraction and refinement workflow used in the proposed system.

Once the candidate masks are collected, the system filters them according to the relative area. A candidate is disqualified if it includes less than 0.2% or more than 35% of the entire picture. The filtering is intended to exclude very small fragments and very large regions that are unlikely to represent useful localized lesion regions. The remaining candidates are scored according to their closeness to the target area fraction of 5% and a boundary-complexity term. The lesion mask with the highest score is chosen as the primary lesion mask. If no valid candidate survives, the implementation uses a fallback selection path.

The minimum and maximum x and y coordinates of the foreground pixels are calculated to determine the bounding box. This is then padded with ~5% of the larger dimension of the crop. This padding is essential, as a lesion-only crop can be too tightly cropped, and can eliminate the surrounding context that is needed for interpretation of the lesion. The resulting processing stage generates a number of outputs, such as the binary mask, an overlay, the crop and the masked crop. These intermediate steps can also be helpful when deploying as they help the operator check if a plausible region has been selected by the segmentation stage.

A key distinction between candidate segmentation and ground-truth lesion segmentation is that the former relies on an algorithm to generate the set of candidates, while the latter depends on annotations provided by experts. No manually annotated lesion data set or accuracy measure (e.g. intersection over union against expert / gold mask) is reported. Hence, the FastSAM stage is considered to be an automatic region proposal process. The following classifiers give an idea of whether or not the resulting representation is useful in classification but do not prove that the masks correspond exactly to biologically annotated lesions.

B. Lesion Refinement in Lab Color Space

The chosen crop is further refined prior to being utilized by MobileViT. The implementation normalizes the crop into the Lab color space, and extracts the luminance channel and the two chromatic channels. The luminance part is subjected to contrast limited adaptive histogram equalization (CLAHE). This operation is designed to make an improvement in local contrast, without merely making greater differences in the overall brightness. Two Chromatic channels have been analyzed by using robust deviations from their median values. Unusual strong pixels do not dominate the estimated background statistics by using median absolute deviation instead of standard deviation.

The implementation is a mixture of the normalized chromatic deviations and a small component of the local luminance enhancement. The response is then scaled to an 8-bit range and thresholded by Otsu's method. After performing morphological opening to remove small isolated regions, morphological closing is used to merge adjacent areas and to close small holes. The refined luminance channel is corrected across with the original chromatic channels to produce the refined crop. The method is fairly conservative as it seeks to alter the luminance data without changing the underlying chromatic structure in areas determined to be locally distinctive.

The MobileViT branch is the last lesion-focused input, which is a refined crop. This operation will result in a lesion branch which will be different from just resizing the FastSAM bounding box. The goal is to improve the existing visual evidence in the candidate region and to minimize small artifacts. Meanwhile, the method has a potential source of error: If FastSAM chooses an inappropriate region, the refinement process can improve the incorrect contents. The quality of the lesions is thus a major drawback of the overall pipeline that is addressed in the present error analysis.

C. ConvNeXt Full-Leaf Branch

We adopt the branch of feature learning using ConvNeXt Tiny with pretrained initialization [9]. The ConvNeXt branch receives the original leaf image rather than a lesion crop. The choice to keep the entire image is based on the fact that identity of the disease may be based on the appearance of more than a single lesion. Classes can be defined by the distribution of the symptoms, the correlation between affected and unaffected areas, the shape of the leaves or overall color distribution. These cues can be maintained if a full leaf representation is used.

The ConvNeXt baseline is trained independently and its performance is interpreted as the contribution of the global information alone. The ConvNeXt baseline is trained with AdamW with a learning rate of 0.0001, weight decay of 0.0001, and gradient clipping with 12 epochs. The pretrained initialization is used to get a good start on the representation learned from the large-scale visual data, and the output layer is fine-tuned to 41 disease classes.

ConvNeXt is not considered to be an ideal, or generally better, architecture. Its role in this study is methodological: it serves as the global pathway in the proposed fusion system. The last comparison is not meant to suggest that ConvNeXt is superior to all the available crop disease models. On the contrary, it claims that the proposed combination of ConvNeXt and lesion-specific MobileViT outperforms the ConvNeXt baseline under the same experimental settings.

D. MobileViT Lesion-Focused Branch

The second branch is MobileViT Small using pretrained initialization [10]. It gets the improved lesion-focused crop instead of the whole leaf. MobileViT is suitable for this branch because its architecture combines convolutional local processing with transformer-style contextual modeling. The study uses cosine annealing for 18 epochs, balanced sampling, label smoothing of 0.1, a learning rate of 0.0003, a weight decay of 0.0001, and gradient clipping.

This is not an attempt to create the complete classifier in a different architecture. It deliberately changes the information presented to the model. The branch is trained to learn from a localized area that has been selected and refined because it is expected to contain disease evidence. This will allow for a controlled test to see if lesion information is useful when removed from the global branch.

The standalone Mobilevit performance is worse than the ConvNeXt performance. This is a useful result and not a missed goal of the research. The information about the overall leaf, the distribution of symptoms or cues that differentiate visually similar classes may be lost in a lesion crop. Although lesion-focused information alone is insufficient, the fact that the fusion model later outperforms ConvNeXt suggests that the lesion branch still provides useful complementary information.

E. Attention-Guided Feature Fusion

Let F_1 be the feature vector extracted by ConvNeXt from the original image, and let F_2 be the feature vector extracted by MobileViT from the “refined region”. First, a 768-dimensional ConvNeXt feature and a 640-dimensional MobileViT feature are used in the implementation of the study. These are both projected into a common 256-dimensional representation via a learned linear projection. This common representation is required because the two backbones do not naturally produce feature vectors of the same dimension.

The projected features are concatenated and passed through a small gating network. The gate consists of a linear layer, a ReLU activation and a final linear layer that yields 2 gate logits. These logits are then used by a softmax operation to produce two non-negative weights that add to 1. The fused representation is computed as a weighted combination of the two projected feature vectors. In compact form, the fusion operation is expressed as:

$$F = \alpha_1 F_1 + \alpha_2 F_2,$$

where F is the fused representation, F_1 and F_2 are the projected ConvNeXt and MobileViT features, and α_1 and α_2 are the learned branch weights. Because the weights are computed from the concatenated features, the model can adapt its preference according to the input rather than applying a constant coefficient. This is the central architectural mechanism of the proposed system.

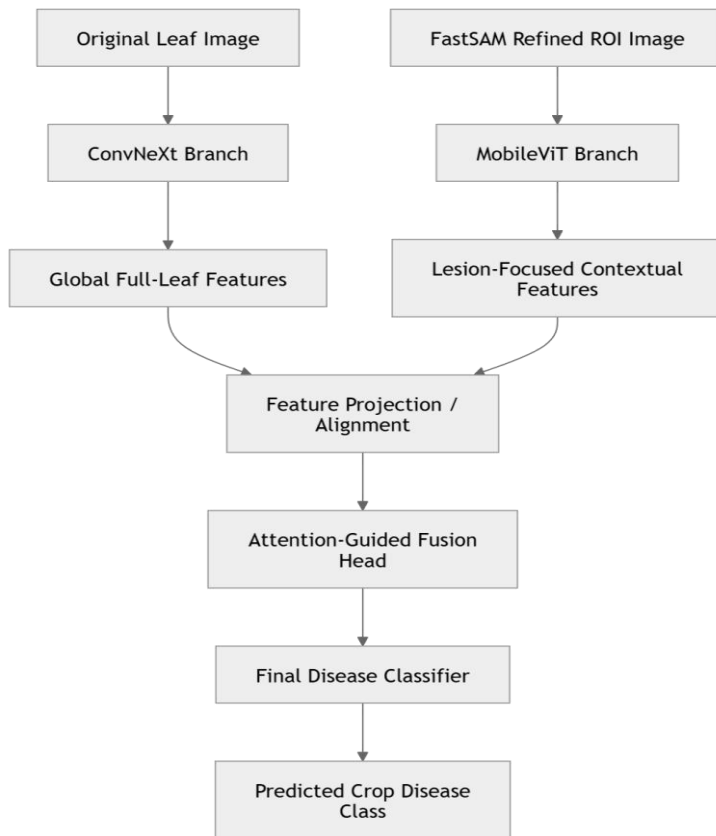


Figure 2: Overall architecture of the proposed Attention-Fused ConvNeXt–MobileViT crop disease detection framework

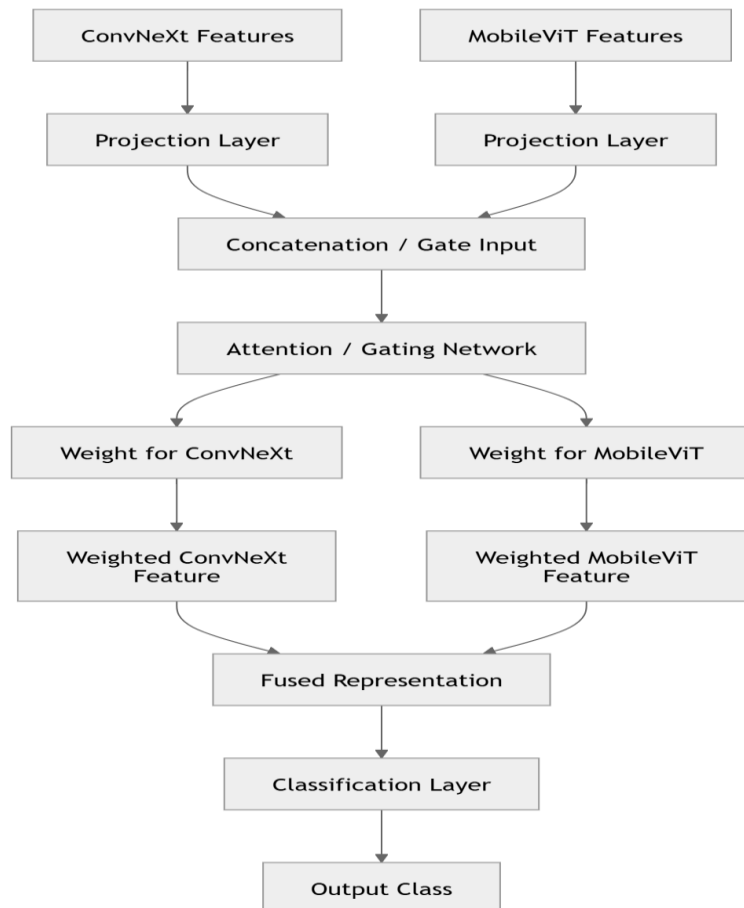


Figure 3: Attention-guided fusion mechanism for combining the ConvNeXt and MobileViT representations

The attention mechanism is related conceptually to the broader attention principle introduced by Vaswani et al. [25], but the present implementation is not a multi-head self-attention transformer block. It is a compact two-branch gating module. This distinction is important for technical precision. The model learns branch importance rather than token-to-token attention over the image.

A potential problem with learned gating is branch collapse, where one pathway receives almost all of the weight and the other becomes functionally irrelevant. To reduce this risk, the fusion training objective includes entropy regularization. If the gate produces probabilities α_1 and α_2 , the entropy term is computed as $H(\alpha) = -\sum_i \alpha_i \log(\alpha_i + \epsilon)$. The implementation subtracts a weighted entropy term from the cross-entropy loss, encouraging a non-degenerate distribution of branch weights during training. The study uses a coefficient of 0.05.

The fusion model is trained in two stages. In the first stage, the backbones are frozen and the fusion head is optimized for five epochs with AdamW at a learning rate of 0.0003. In the second stage, selected final backbone stages are unfrozen and jointly fine-tuned for another five epochs at a lower learning rate of 0.00003. This staged procedure allows the gate and classifier to first learn how to combine stable feature representations before introducing limited adaptation of the underlying backbones.

F. End-to-End Training Pipeline

The complete training pipeline can be understood as a sequence of linked operations. First, the dataset is normalized and split into training, validation, and testing partitions. Second, FastSAM is applied to generate lesion candidates for the images that require lesion-focused processing. Third, candidate masks are filtered and the primary region is selected. Fourth, the crop is refined in Lab space and stored as the paired lesion image. Fifth, the ConvNeXt and MobileViT branches are trained independently to establish baselines. Finally, the two trained feature streams are connected to the projection, gate, and classification layers and optimized through the staged fusion procedure.

The same canonical class mapping is used throughout the experiment. Model checkpoints, training histories, classification reports, confusion matrices, and class mappings are saved to persistent storage. This is relevant to reproducibility because the evaluation depends on the exact relationship between class indices and disease names. The study implementation also supports interrupted training and later evaluation by saving intermediate artifacts.

Table 2: Main experimental configurations reported in the study

Model	Input	Resolution	Main configuration
ConvNeXt Tiny	Original leaf	224×224	12 epochs; AdamW; LR 1e-4; WD 1e-4; gradient clipping
MobileViT Small	Refined ROI	288×288	18 epochs; balanced sampling; label smoothing 0.1; LR 3e-4; cosine annealing
Fusion	Original + refined ROI	224×224 + 288×288	5+5 epochs; frozen then partial fine-tuning; entropy $\lambda=0.05$

Evaluation Metrics and Experimental Protocol

A. Accuracy

The number of test samples correctly classified, divided by the total number of test samples. It shows a direct assessment of the overall classification correctness, and can mask misclassification on classes with fewer instances. In this experiment, the test set consists of 635 images from 41 classes, with a different class support. In addition to macro-level precision, recall and F1-score, accuracy is reported.



B. Precision, Recall, and F1-Score

Precision is the percentage of predictions that fall into the correct class, while recall is the percentage of the class that is found by the model. F1 is the harmonic mean of precision and recall. Macro averaging calculates an independent metric for each class, and then averages 41 times. Instead, weighted averaging weights each class according to its level of support. The study focuses on macro F1 since it offers a class-balanced perspective of performance and is not so skewed towards the highest categories.

C. ROC-AUC

This study also compares ROC curves and macro-average Area under ROC (mAUC) for one vs rest. In each class the target class is considered positive and all other classes as negative. AUC is a summary of model's performance when it is used to sort positive examples to the top of the ranking list above negative examples for various classification thresholds. The macro-average AUC results are presented for ConvNeXt (0.9951), MobileViT (0.9713), and Fusion (0.9978).

D. Confusion-Matrix Analysis

Confusion matrices offer a more detailed analysis of the behavior of classes, highlighting the focus and errors of the predictions. This study has a total of three matrices for ConvNeXt, MobileViT, and Fusion. An important audit observation is that these matrices are not 41×41 matrices of disease classes, but are matrices of crop categories displayed. Six groups are shown: Cotton, Rice, Maize, Wheat, Sugarcane, General/Pests. They are thus largely a summary of the error structure of the crop level, and should not be considered as a full 41 class confusion matrix. The distinction is maintained here in the present paper, so as not to overstate what is shown by the figure.

E. Fairness of the Internal Comparison

The same held-out test set and the same class mapping is used to evaluate all three models. This is to control one of the major components of experimental variation. However, the models are not given the same input representation: ConvNeXt is provided the original leaf; MobileViT is provided with a refined region of the lesion; and Fusion is given both the original leaf and the refined region of the lesion. This is done on purpose as the goal is to compare the various information pathways, not just different classifiers applied to the same pixels.

EXPERIMENTAL RESULTS

A. Overall Test Performance

The final evaluation demonstrates a clear performance hierarchy. ConvNeXt achieved 0.8976 test accuracy, 0.8394 macro precision, 0.8089 macro recall, 0.8155 macro F1, and 0.8921 weighted F1. MobileViT achieved 0.7276 accuracy, 0.6992 macro precision, 0.6761 macro recall, 0.6807 macro F1, and 0.7271 weighted F1. The Fusion Model achieved 0.9307 accuracy, 0.8697 macro precision, 0.8479 macro recall, 0.8527 macro F1, and 0.9288 weighted F1.

Table 3 : Test performance of ConvNeXt, MobileViT, and the proposed Fusion Model

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
ConvNeXt	0.8976	0.8394	0.8089	0.8155	0.8921
MobileViT	0.7276	0.6992	0.6761	0.6807	0.7271
Fusion	0.9307	0.8697	0.8479	0.8527	0.9288

The Fusion Model improves accuracy by 3.31 percentage points over ConvNeXt and by 20.31 percentage points over MobileViT. Its macro F1 improvement over ConvNeXt is 3.72 percentage points, while the

improvement over MobileViT is 17.20 percentage points. The simultaneous increase in accuracy and macro F1 is important because it indicates that the gain is not limited to the majority-supported classes. The macro metrics also increase, showing that the fusion representation provides a more balanced classification result under the evaluation protocol.

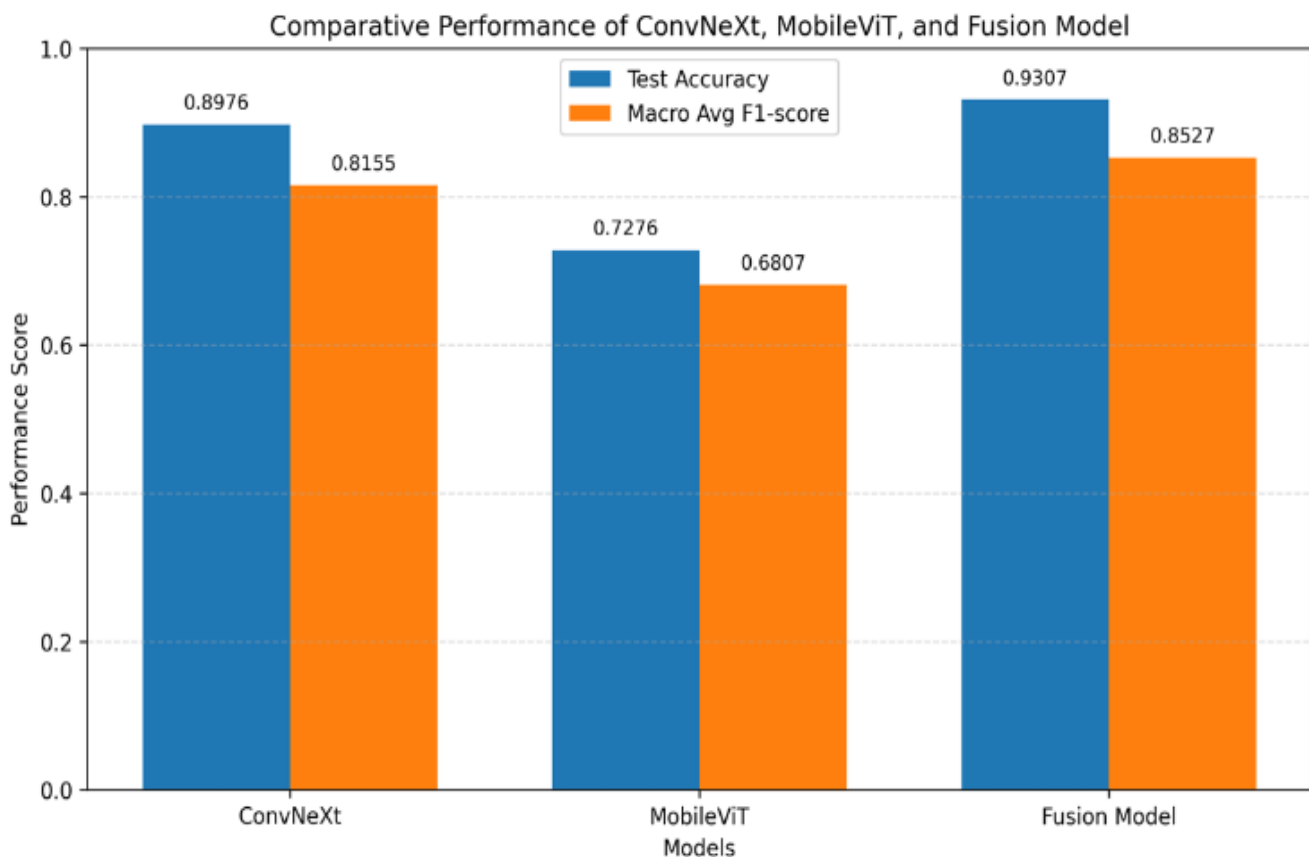


Figure 4 : Comparative test performance of ConvNeXt, MobileViT, and Fusion based on accuracy and macro F1-score

The ConvNeXt result demonstrates that the complete leaf already contains substantial disease information. This is consistent with the role assigned to the global branch: the model can use the spatial arrangement and broader appearance of symptoms rather than depending only on a small lesion region. The result also establishes a strong baseline, making the additional gain from fusion more meaningful than if the global branch were weak.

The MobileViT result is substantially lower. This should not be interpreted as evidence that lesion information is useless. The lesion branch has been deliberately constrained to a refined ROI, which can remove disease context as well as irrelevant background. A lesion can be visually distinctive while still requiring information about its location, distribution, surrounding tissue, or crop identity. The standalone result therefore demonstrates a limitation of localized evidence when it is used without the global pathway.

The Fusion result indicates that the two branches contain complementary information. The model does not simply choose the stronger branch in every case because the gate is input dependent. Instead, it can combine full-leaf context with localized lesion evidence. The improvement over ConvNeXt suggests that at least some test samples contain discriminative information in the lesion-focused representation that is not fully captured by the original image pathway.

B. ROC-AUC Results

A second perspective of the model discrimination is given by the ROC analysis. The Fusion Model obtained a macro average AUC of 0.9978, ConvNeXt got 0.9951, and MobileViT got 0.9713. The three values are

significantly greater than the random reference value (0.50). The study states that the Fusion curve is the closest to the top left, while ConvNeXt is the second closest, and then MobileViT.

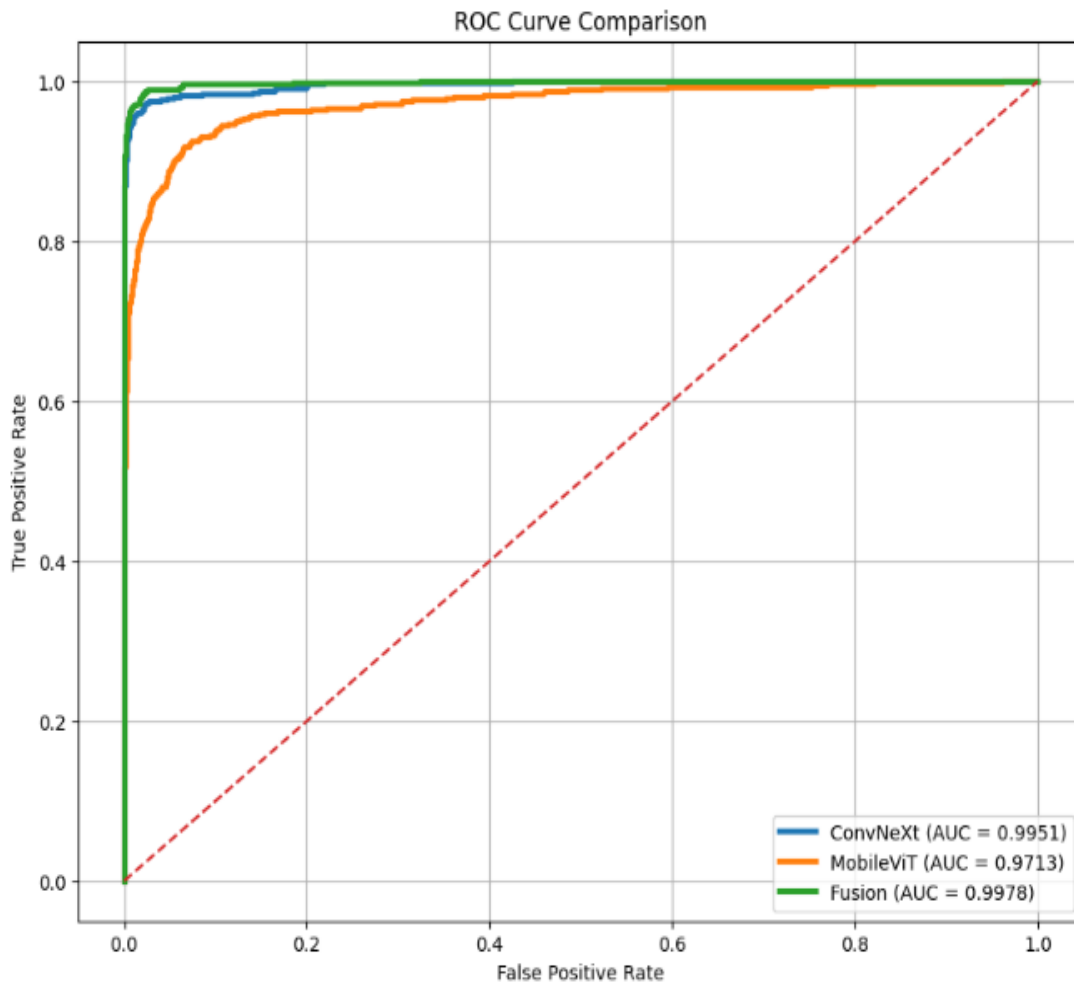


Figure 5: ROC curve comparison for ConvNeXt, MobileViT, and Fusion Model

The accuracy difference between Fusion and ConvNeXt is greater than the AUC difference between these two models, which is small. AUC measures the quality of the rankings at different thresholds, and accuracy is computed at the model's class decision threshold. However, a high AUC does not necessarily mean that every class is equally well classified at a single decision threshold. The higher AUCs of the Fusion model further demonstrate that the fusion model is discriminative in this experiment, complementing its high accuracy and macro F1 score.

The MobileViT AUC of 0.9713 is still high, although it has a lower classification accuracy. Therefore, while the final top-1 predictions of the lesion-based model are less reliable than those of the global and fused models, the lesion-based model has some class information that is useful. This observation is consistent with the main hypothesis of the study, which states that the lesion branch is not necessarily the optimal individual classifier to help the fusion system.

C. Grouped Confusion-Matrix Analysis

Each model has a crop-grouped confusion analysis in the study. The grouped visualization cannot display a full matrix of classes, as the original experiment has 41 classes. The rows are groups of actual crops and the columns are groups of predicted crops. The six groups include the following: Cotton, Rice, Maize, Wheat, Sugarcane, and General/Pests. The matrix values are row-normalized for the visualization; the cell annotations are counts. Both the raw support and the relative row behavior can be checked in this combination.

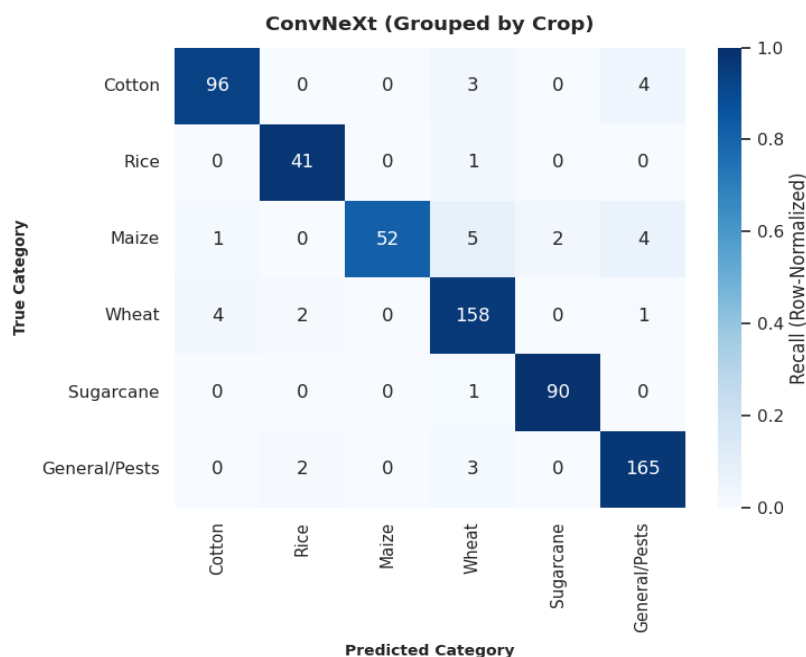


Figure 6 : Crop-grouped confusion matrix for the ConvNeXt Model

ConvNeXt makes 96 correct predictions on the Cotton row, with fewer predictions for Wheat and General/Pests. Rice has 41 correct predictions and 1 prediction that points to Wheat. Maize has 52 correct predictions, the same for Wheat, 2 for Sugarcane and 4 for General/Pests. Wheat has 158 correct predictions and a small spillover in the Cotton, Rice and General/Pests categories. Sugarcane has made 90 correct predictions and 1 prediction for Wheat. General/Pests has 165 correct predictions, a few of which are given to Rice and Wheat. The values indicate that the global model is mostly consistent at the crop-group level, while still having some confusion across crop-groups.

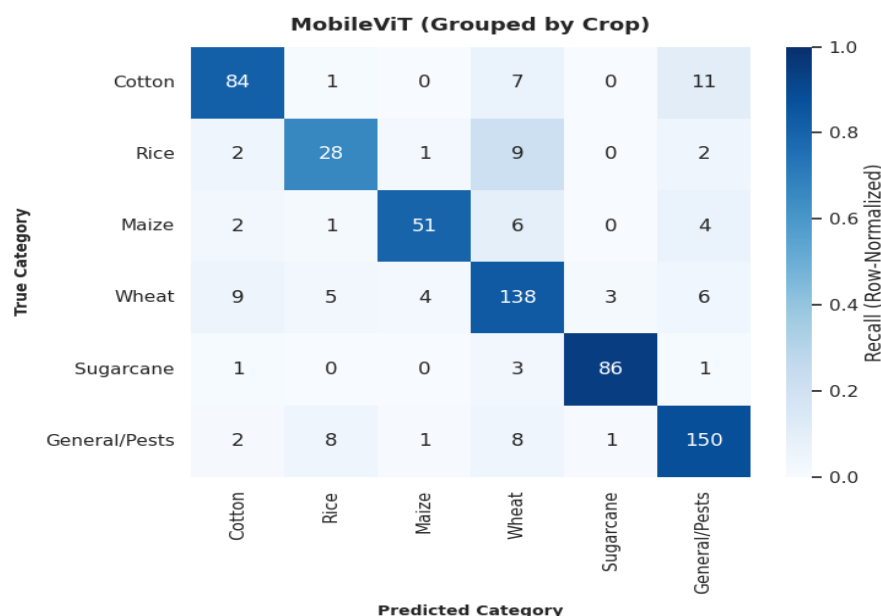


Figure 7: Crop-grouped confusion matrix for the MobileViT Model

The grouped MobileViT matrix contains more off-diagonal errors. For Cotton, there are 84 correct predictions with 7 general predictions for Wheat and 11 for Pests. The model correctly classifies 28 Rice samples and 9 Wheat samples. Maize has 51 correct predictions and smaller errors towards Cotton, Rice, Wheat and General/Pests. Wheat has 138 correct predictions and there were errors across Cotton, Rice, Maize, Sugarcane, and General/Pests. Sugarcane has 86 correct predictions and four non-Sugarcane predictions. General/Pests

makes 150 correct predictions and has a few predictions assigned to Rice and Wheat. This visual pattern is similar to the lower standalone MobileViT scores.

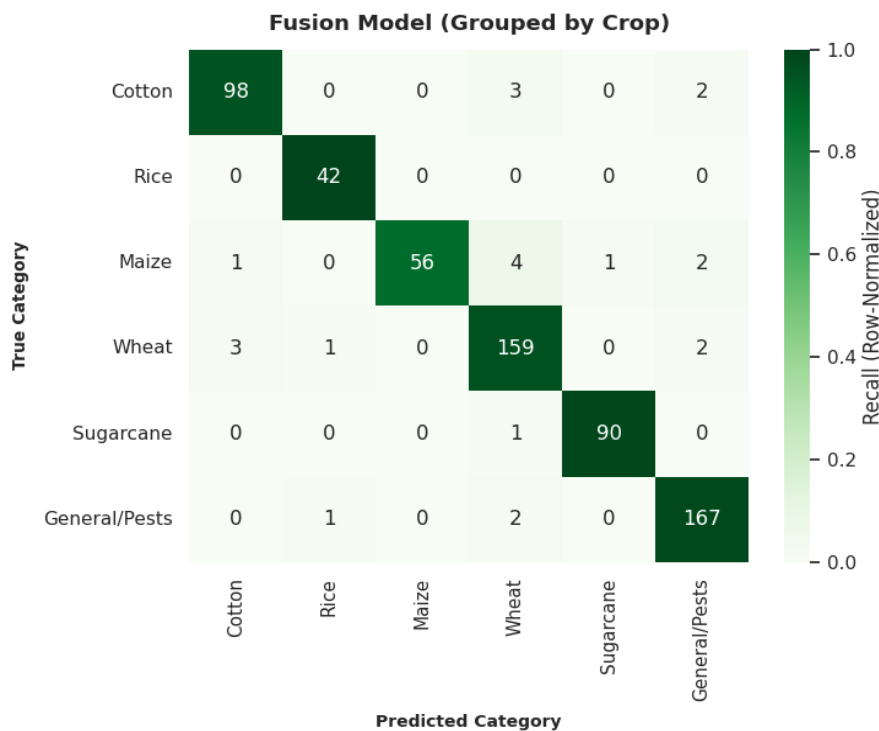


Figure 8 : Crop-grouped confusion matrix for the proposed Fusion Model

The Fusion matrix is much more diagonalized. In the case of cotton, there are 98 correct predictions, and a few predictions for Wheat or General/Pests. Rice has 42 correct predictions and no cross-group errors are shown. Maize has 56 correct predictions and small spillover to Cotton, Wheat, Sugarcane and General/Pests. There are 159 correct predictions in wheat, and a few predictions are given elsewhere. In sugarcane, 90 predictions are correct and 1 prediction is made for Wheat. There are 167 correct predictions in the General/Pests category, 1 prediction for the Rice category and 2 predictions for the Wheat category. The grouped visualization thus directly qualifies the numerical finding that the fusion system leads to a more consistent classification pattern.

The confusion analysis also explains why aggregate accuracy is not necessarily a measure of perfect recognition. Even for the Fusion Model, there are non-zero off-diagonal elements. For some disease categories, the lesions may be colored, textured, severe, or spatial similar; some pest categories may be similar in appearance to disease symptoms. Further, the grouped matrix will not indicate which particular disease labels caused each mistake. For that analysis, it would need a 41×41 class level matrix. The present paper is not a reconstruction of unsupported class level counts, but reports only the matrix available in the study, as this was intentional.

The grouped matrices will allow a wider range of interpretation of branch behavior. The full-leaf ConvNeXt representation gives more context in terms of whole crop than the lesion-only MobileViT pathway. The lesion-focused pathway, on the other hand, provides extra local information that can be used by the fusion model to decrease some of the errors detected in the global branch. This is an example of what the architectural goal of complementary evidence looks like.

D. Comparison with Selected Literature

The paper compares the proposed system to some crop disease classification studies. While these comparisons can be helpful for contextualization, they should not be used as controlled head-to-head comparisons, since the data used for the studies, the number of classes, the crop grown, the preprocessing procedure, and the evaluation protocol are different. This paper, thus, describes the literature comparison.

Table 4: Selected literature comparison preserved from the study context. Numerical comparison is limited by differences in datasets and protocols

Study	Approach	Task	Reported result in study
Zhu et al. [1]	Multiscale CNN + ViT	Plant disease	99.86%, 97.50%
Tonmoy et al. [2]	MobilePlantViT	Generalized plant disease	Reported in study literature review
Rodrigo et al. [3]	MobileViT-based classifier	Rice leaf disease	93.14%, 99.6%
Guo [4]	CNN	Crop disease images	Referenced as CNN baseline literature
Ergün [5]	ConvNeXt + ViT fusion	Mango leaf disease	99% plus
Borhani et al. [15]	Vision Transformer	Plant disease	Referenced as ViT approach
Aboelenin et al. [16]	CNN + ViT	Plant leaf disease	Referenced as hybrid framework
Huang et al. [19]	ConvNeXt + Transformer	Apple leaf disease	Referenced as hybrid model
Proposed	ConvNeXt + MobileViT + FastSAM	41-class crop disease	93.07%

The primary scientific interest in the comparison is not that the proposed model is the most accurate numerically of all the studies published. This would be incorrect since the experiments are not comparable. However, the literature indicates that hybrid feature learning, light transformer models, convolutional models and segmentation-assisted models are all ongoing directions. In the present study, these ideas are combined in a single implemented pipeline and the internal contribution of each component is assessed.

E. Error Analysis

The study points out that there are multiple types of errors. The first issue is that the lesion-focused branch relies heavily on the quality of FastSAM's candidate mask. When the region is missing or contains irrelevant tissue or does not match strongly with the disease symptom, the MobileViT branch is presented with a degraded representation. The classification pipeline now includes the segmentation quality as a preprocessing detail, since the model is trained with these generated regions.

Second, the lesion crop may remove useful contextual information. Symptoms may not be confined to one area of the leaf, but spread throughout the leaf and this may be the basis of a disease description. This may help explain why MobileViT performs worse than ConvNeXt when used alone. The fusion architecture overcomes this limitation by keeping the entire image in the ConvNeXt branch.

Thirdly, the gate itself may become unbalanced. To prevent all forms of branch dominance, and to encourage more meaningful use of both pathways, entropy regularization is added. In fact, it is not necessarily desirable for the two branches to receive equal weights, since there may be legitimate differences in the quantities of useful global and local information that the two inputs carry. Its goal is to adaptively contribute, not to force symmetry.

Fourth, the data source is curated. Shadows, clutter, multiple leaves, occlusion, illumination changes and other variations are noted explicitly as being caused by field conditions. A model can thus perform well on a subset of the test set without being as effective in an uncontrolled agricultural setup. This restriction is significant for a disease-diagnosis application since the user can take pictures of leaves at any angle, distance and light source.

F. Interpretation of the Fusion Gain

One of the key findings of the study is the improvement in accuracy from 89.76% to 93.07% for ConvNeXt and Fusion, respectively. The absolute percentage gain is relatively small when compared to the difference between ConvNeXt and MobileViT, but it is significant as ConvNeXt is already a good baseline. The fusion system thus does not take advantage of a poor classifier, but rather refines a good global representation with another pathway focused on the lesion.

This macro F1 improvement of 81.55% to 85.27% is particularly significant. A model may be stronger on classes with high test support and weaker on minority classes and, as a result, more accurate. The macro F1 result proves that the improvement in fusion is also seen in class balanced performance. This is not only because of its informativeness, but also because the class support is different for both classes.

The results also show that the optimal component test architecture may not be optimal for the entire system architecture. MobileViT is less powerful as a single entity, but it's still effective when combined with other entities. A typical paradigm in multi-branch representation learning is that a branch could provide complementary characteristics which are partly redundant or inadequate for individual classification, but can be helpful when joined with other representation. The present experiment supports that interpretation in the context of the present data set and protocol.

Deployment Prototype

A. Deployment Architecture

The study extends the research into a Streamlit-based interactive prototype. The prototype supports three inference modes: ConvNeXt, MobileViT, and Fusion. ConvNeXt uses the original uploaded image, MobileViT uses FastSAM-based lesion preprocessing, and Fusion uses both the original image and the refined lesion crop. This mirrors the training-time information flow and reduces the risk of implementing a deployment path that differs from the evaluated model.

The deployment implementation is modular. The model-loading component retrieves the selected trained checkpoint and class mapping. The FastSAM inference module performs segmentation with retry logic, evaluates candidate masks, generates the standard outputs, and performs lesion refinement. The prediction component applies the appropriate preprocessing and returns top-k class probabilities. The Grad-CAM utility registers hook on the selected target layer, captures activations and gradients, and produces a normalized class activation map. For the fusion mode, branch weights can also be exposed so that the user can inspect the relative contribution of the two streams.

The system is therefore more than a single prediction function. It represents an end-to-end workflow in which image acquisition, preprocessing, inference, and interpretation are connected. The study also describes automatic checkpoint setup support so that required FastSAM resources can be prepared when the prototype is executed in a new environment.

B. Functional Components

Table 5 : Functional components of the deployment prototype.

Component	Function
Image upload	Accepts a new leaf image for inference
Model selection	Selects ConvNeXt, MobileViT, or Fusion
FastSAM preprocessing	Generates lesion-focused input for relevant modes
Prediction	Returns disease class and confidence information
Top-k output	Displays ranked prediction probabilities



Grad-CAM	Provides visual evidence associated with model output
Fusion inspection	Displays branch weights and branch-wise interpretation

Crop Disease Detection Prototype

Attention-Fused ConvNeXt and MobileViT Models with FastSAM Lesion Extraction for Crop Disease Detection

Select Model

ConvNeXt

MobileViT

Fusion

FastSAM weights already available: FastSAM-x.pt

Model ConvNeXt	Device CPU	Checkpoint Folder baseline_convnext	FastSAM Weights FastSAM-x.pt
--------------------------	----------------------	---	--

Upload a leaf image

 Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

Browse files

Upload a leaf image to run preprocessing, prediction, and Grad-CAM.

Figure 9: Example deployment output with lesion-focused preprocessing and visual explanation

The deployment demonstration is important because it tests whether the research workflow can be executed interactively rather than only through training scripts. The prototype accepts an image, internally performs the required preprocessing, loads the selected model, runs inference, and presents the result. This provides a practical proof of concept for the research architecture.

However, the prototype should not be described as a validated agricultural decision-support product. The study does not report a prospective field trial, user study, latency benchmark on representative mobile devices, or agronomist-validated diagnostic accuracy. It therefore demonstrates functional feasibility rather than real-world clinical or agricultural validity. This distinction is retained in the final paper.

DISCUSSION

Results from metrics, ROC analysis and the grouped confusion matrices presented support the central hypothesis of the study. The accuracy of the Full-leaf Information is very high at 89.76%, which is very informative. The information from lesions is not enough on its own, with a MobileViT accuracy of 72.76%. However, the information in the lesion pathway is complementary since after attention gated fusion, the final accuracy is 93.07% and the macro F1 is 85.27%.

One view is that the pathways observe different levels of disease organization, with global and local pathways. The global branch can identify the overall appearance of leaves and the distribution of symptoms across the leaf. The lesion branch can enhance local patterns that can be quite small and thus are diluted in a full-leaf representation. The fusion gate combines these representations with input-dependent weights. The architecture therefore considers disease recognition as a problem of multi-scale representation, instead of a single-scale classification problem.

A second key point is that the extraction of lesions should not be considered just a convenience in the pre-processing steps. It alters the information that is available for classification. A good ROI can make one symptom more visible relative to other ones, and a bad ROI can either take away helpful context, or add unnecessary pixels. The study therefore does not assume that the first segmentation output is always correct, but relies on FastSAM and heuristic selection and refinement instead. The changes in this stage will directly affect the MobileViT branch and subsequently the fusion result, which should be improved in the future.

It is also important to mention the effect of entropy regularisation. A robust ConvNeXt branch may win the gate just by being easier to optimize if there is no mechanism that would try to prevent early collapse. During fusion training, the entropy term encourages meaningful use of both pathways. The value of the coefficient 0.05 is part of the experimental setup. This study does not present any separate ablation study with a different value of this coefficient, hence 0.05 is not claimed to be an optimum value.

Explanations must also take the class structure into consideration. There are 41 classes in the experiment comprising of various crops and disease/pest categories. This is a more diverse environment than one of a single crop disease benchmark. Meanwhile, the use of a selected subset of 4000 images reduces the number of images used to evaluate the system. The reported performance is for the controlled subset and split, and not for the whole source dataset.

The literature comparison indicates that the proposed architecture is in line with the general development in hybrid agricultural vision research [1], [2], [5], [16], [19]. It also relates to previous contributions on convolutional feature learning [7, 13], efficient architectures [14, 23], attention mechanism for image representation [25] and transformer architectures. The employment of Grad-CAM [12] and a deployment prototype [21] also puts the study in the context of developing intelligently explainable and usable agricultural systems.

Finally, there are methodological weaknesses that should be considered in interpreting the study. No external validation data is provided. A statistical significance test across repeated independent runs is not reported. No expert-annotated benchmark for segmentation of lesions is reported. There is no formal latency or memory benchmark for deployment. The results therefore provide useful internal experimental findings and do not provide evidence of universal superiority, robustness in the field, segmentation accuracy or production readiness.

The limitations do not detract from the contribution, they just set boundaries for it. The contribution is a fully implemented framework that shows the benefit of merging full-leaf and lesion-based information within a controlled 41 class experiment. The next step in this research is to evaluate the same gain when the model is applied to independent field data, and to more varied acquisition conditions and stronger lesion annotations.

Reproducibility and Implementation Details

The study implementation includes sufficient procedural information to reproduce the central architecture. The fixed random seed applied in the dataset preparation is 42, the target number of images is 4000, the class cap is 200, and the training to validation split is 70:15. FastSAM uses FastSAM-x weights and a sequence of re-tries (confidence and IoU). Candidate masks are transformed to Boolean arrays and mask area fraction is used to filter. The area fraction target is 0.05 and the mask score consists of an area deviation term as well as a boundary-complexity term. The implementation discussion is also consistent with the general discussion of deep learning that is described by Goodfellow, Bengio, and Courville [22].

The fused model has a feature dimension of 768, a feature dimension of 640 for MobileViT, a projection dimension of 256, and 41 classes for output. The concatenated 512 dimensional projected representation is passed through a 256 units hidden layer to two gate logits in the gating network. These logits are then converted to branch weights by the Softmax. The fused 256 dimensional representation is fed to a linear 41-class classifier. The details are more precise than saying that the models are simply concatenated.

The implementation of lesion refinement is based on CLAHE, clip limit 2.0, and tile grid size of 8×8. Statistics of median and median absolute deviation are calculated for the chromatic channels. Thresholding of the resulting deviation map is performed by the Otsu's method, followed by morphological processing using an elliptical kernel derived from the image dimensions. These implementation details are important as the exact refinement process is used to determine the lesion branch.

Checkpoints, histories, classification reports, confusion matrices and class mappings are saved within the training workflow. This is especially useful for the paper-based evaluation because the reported results are

based on saved evaluation artifacts rather than manual transcription from a single run. The deployment prototype also uses the same logic for lesion preprocessing for consistency in training and inference.

LIMITATIONS AND THREATS TO VALIDITY

The first major threat to external validity is dataset limitation. The study is based on a public dataset and a controlled set of 4000 images. The images are not necessarily representative of all field conditions, although the 41-class setting is diverse. Images may contain multiple leaves, the background can be complex, images may contain shadows, varying focus, and symptoms at different stages of development, all within a natural image. When the trained classifiers are used on images other than the training set, they may encounter distribution shift.

The second one is related to class support. The study clearly states that class support is not the same so there is an emphasis on macro metrics. The grouped confusion matrices also show that the crop categories are not of the same size when tested. This difference in class support should be considered when interpreting the grouped confusion matrices. The use of macro F1 does not solve the problem of the sampling limitation of large classes, but it does alleviate the impact that these large classes have on the overall score.

Third, lesion localization, which is also difficult. This is an automated candidate generator, not a ground-truth disease lesion annotator, and is called FastSAM. The segmentation precision, recall, IoU and Dice score are not provided in the study against expert masks. Currently, it is not feasible to distinguish between classification errors due to the misclassification of the backbone and classification errors due to misclassification of the lesions.

The fourth restriction is that not all components have been reported as being ablatable. In the paper, the two complete baselines and the final fusion model are compared; however, there are no controlled experiments conducted to test the effects of only the Lab refinement, only the mask-ranking strategy, or only the entropy coefficient. The overall design is supported by the observed fusion gain, although the distinctiveness of each single preprocessing option is not proven.

The fifth restriction is the deployment evaluation. It is worth noting that the prototype does not measure the time required to infer, memory usage, concurrent users, mobile device performance, or even perform a field usability study. The deployment contribution should, therefore, be called a prototype of a research application of agriculture rather than a production grade agricultural application.

Last, the paper does not assert that a diagnosis is always reliable. The results of the model should be considered an automatic image-classification result in the evaluated class set. It is not a replacement for agricultural knowledge and experience; it should not be used when the image is not of a good quality and it should not be used when images are from a disease class not included in the class labels in the training data.

Future Research Directions

In future, the localization of lesions will be improved first. The present heuristic chooses one of the main candidates from the FastSAM outputs. A more robust system would compare several candidate regions, feature an ROI selection method that would be class-aware, or utilize expert annotations to train and/or test a disease-specific segmentation module. Two possible directions for development were identified in the study: improved mask ranking, and alternative segmentation methods, as well as refinement by feedback.

Second, greater validation is needed. The model needs to be tested on out-of-scope data and on field images with natural illumination, background and disease severity and crop presentation to assess its performance. A split from a different source will give a less good estimate of generalization than will a cross-dataset split.

Thirdly, the fusion mechanism can be investigated in greater detail. Further experiments can be conducted to investigate the distribution of gate weights within each class of disease, to explore whether some classes are more heavily dependent on local or global evidence and to examine the impact of various entropy coefficients.



This would make the gate not just a mechanism for performance but a representation of branch preference that may be interpreted.

Fourth, explainability can be investigated further. Comparisons can be made between grad-cam maps and lesion masks, and correctly classified and misclassified examples can be analyzed to see if the model is targeting plausible disease areas. The current implementation already supports branch-wise Grad-CAM analysis [12].

Fifth, deployment can be optimized. The prototype can be extended to become an actual API or even a cellular app, with compression of models and efficient inferences. This is not to be described in a merely qualitative way, but should be accompanied by latency, memory and field usability measurements. This is relevant in the context of efficient vision models given in the literature [10, 14, 23].

Last but not least, uncertainty-aware prediction and open-set behavior can be explored in future research. The present 41-class classifier takes it for granted that the correct label is in the given set of known labels. An ideal agricultural system should also identify images that do not belong to the specified categories or those of poor quality that are not suitable for a correct prediction. This direction is not being exploited in the current experiment and is therefore not a current capability, but a direction for future work.

CONCLUSION

This paper proposed an attention fusion dual-branch model for multi-class crop disease detection. It combines FastSAM-based lesion candidate extraction and refinement, ConvNeXt-based full-leaf representation learning, MobileViT-based lesion-focused representation learning, and a learnable attention gate. The design was inspired by the fact that the recognition of the disease may be based on the appearance of the lesions as well as the general context of the leaf.

The controlled experiment was conducted using a subset of 4000 images with 41 classes and a fixed split of 2777 training images, 588 validation images and 635 test images. The ConvNeXt baseline model obtained an accuracy of 89.76% and macro F1 of 81.55%. The MobileViT baseline model, which focuses on lesions, obtained an accuracy of 72.76% and macro F1 of 68.07%. The proposed Fusion Model achieved 93.07% accuracy, 86.97% macro precision, 84.79% macro recall, 85.27% macro F1, and 92.88% weighted F1. The macro-average ROC-AUC value was 0.9978.

Grouped confusion-matrix analysis provides qualitative evidence of differences in crop-level error patterns. In the MobileViT matrix, there are more off-diagonal errors and in the Fusion matrix, more errors are concentrated in the crop-group diagonal. The analysis is stated intentionally grouped by crop and not as a complete 41-class confusion matrix because that is how the results of the analysis are presented in the study. This careful interpretation is to avoid using the figure to make a stronger claim than the evidence allows for.

Overall, the conclusion is that lesion-focused information can be helpful when used in conjunction with full-leaf information. The lower MobileViT performance highlights the risk of losing valuable contextual information when relying only on localized evidence. The good result in the fusion component indicates that local information can gain value if presented together with the global representation. This finding is the main motivation for the proposed attention-fused architecture.

The prototype of deployment shows that the framework can be linked to an interactive image-based workflow. It enables pre-processing, model inference, confidence reporting and Grad-CAM visualization. However, it is still a controlled research experiment, and it is not claimed to be field validated, to be able to segment ground truth accurately or to be production ready. Independent testing in the field, better lesion localization, deeper gate analysis, better evaluation of explainability, and efficient deployment are still key future challenges.

REFERENCES

1. Zhu, D., Tan, J., Wu, C., Yung, K., & Ip, A. W. H. (2023). Crop disease identification by fusing multiscale convolution and vision transformer. *Sensors*, 23(13), 6015.

2. Tonmoy, M. R., Hossain, M. M., Dey, N., & Mridha, M. F. (2025). MobilePlantViT: A mobile-friendly hybrid ViT for generalized plant disease image classification. *arXiv preprint arXiv:2503.16628*.
3. Rodrigo, K. U. K., Marcial, J. H. H. S., Brillo, S. C., Mata, K. E., & Morano, J. C. (2025). An enhancement of CNN algorithm for rice leaf disease image classification in mobile applications. *Journal of Information Systems Engineering and Management*, 10(6s).
4. Guo, J. (2023). Classification of crop disease images based on convolutional neural network. *Advances in Computer and Communications*, 4(3), 205–209.
5. Ergün, E. (2025). Attention-enhanced hybrid deep learning model for robust mango leaf disease classification via ConvNeXt and vision transformer fusion. *Frontiers in Plant Science*, 16, 1638520.
6. Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*.
7. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, M., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
9. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11976–11986).
10. Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
11. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., & Wang, J. (2023). Fast segment anything. *arXiv preprint arXiv:2306.12156*.
12. Selvaraju, R. R., Cogswell, A., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 336–359.
13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).
14. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, R., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1314–1324).
15. Borhani, Y., Khoramdel, J., & Najafi, E. (2022). A deep learning based approach for automated plant disease classification using vision transformer. *Scientific Reports*, 12, 11554.
16. Aboelenin, S., Elbasheer, F. A., Eltoukhy, M. M., El-Hady, W. M., & Hosny, K. M. (2025). A hybrid framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer. *Complex & Intelligent Systems*, 11, 142.
17. Gebremedhin, T. G., Asegede, H. S., Tesheme, B. W., Gebremichael, T. B., & Redae, K. G. (2025). Automated plant disease and pest detection system using hybrid lightweight CNN–MobileViT models for diagnosis of indigenous crops. *arXiv preprint arXiv:2512.11871*.
18. Chen, J., Hu, H., & Yang, J. (2024). Plant leaf disease recognition based on improved SinGAN and improved ResNet34. *Frontiers in Artificial Intelligence*, 7, 1414274.
19. Huang, X., Xu, D., Chen, Y., Zhang, Q., Feng, P., Ma, Y., Dong, Q., & Yu, F. (2025). EConv-ViT: A strongly generalized apple leaf disease classification model based on the fusion of ConvNeXt and transformer. *Information Processing in Agriculture*, 12, 466–477.
20. Ashurov, M., Ochilov, B., Park, S., & Kulekov, E. (2025). Enhancing plant disease detection through deep learning: A depthwise CNN with squeeze and excitation integration and residual skip connections. *Frontiers in Plant Science*, 15, 1505857.
21. Pantelaios, D., Vafeiadis, A., Iosifidis, A., & Tzovaras, D. (2025). AI-driven smart agriculture using hybrid transformer-CNN for real-time disease detection in sustainable farming. *Scientific Reports*, 15, 25817.
22. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.



23. Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
24. Salman, Z., Saeed, A., Asim, M. N., Ibrahim, M. A., Dengel, W., & Ahmed, S. (2025). Plant disease classification in the wild using vision transformers and mixture of experts. *Frontiers in Plant Science*, 16.
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).
26. Ali, J. (2024). 20k+ multi-class crop disease images [Dataset]. Kaggle.
<https://www.kaggle.com/datasets/jawadali1045/20k-multi-class-crop-disease-images>