AI Systems Engineering
https://aise.cultechpub.com/aise
E-ISSN: 3093-8449
Vol. 2, No. 1, October 2026
DOI: 10.64229/1k870c91

Article

MultiBranchFusionNet: A Lightweight Attention-Guided Dual-Branch Framework for Tomato Leaf Disease Classification Using Heterogeneous Field and Public Images

Olayiwola Charles Adesoba[1,*], Waliu Olalekan Apena[1], Folasade Mojisola Dahunsi[1], Ayodele Martins Ajayi[2]

[1]Department of Computer Engineering, Federal University of Technology, Akure, Nigeria

[2]Department of Crop, Soil and Pest Management, Federal University of Technology, Akure, Nigeria

[*]Correspondence author: Olayiwola Charles Adesoba, ocadesoba@futa.edu.ng

Abstract

Tomato foliar diseases reduce yield, degrade fruit quality, and cause major economic losses, especially in regions such as Nigeria where diagnosis still relies largely on manual visual inspection. Although deep learning has improved automated plant disease recognition, many models are trained on controlled datasets that do not capture field variability. This study presents an attention-guided multi-branch transfer learning framework for tomato leaf disease classification using a heterogeneous dataset of 20,245 images from South-West Nigerian farms and public sources, spanning ten disease categories. Field-acquired labels were assigned through visual symptom assessment with expert adjudication of uncertain cases and should therefore be interpreted as symptom-based classifications rather than pathogen-confirmed diagnoses. During preliminary backbone screening, MobileNetV2 was selected as the pretrained feature-extraction pathway because it combined strong recorded classification performance with a compact model profile. MultiBranchFusionNet integrates a frozen ImageNet-pretrained MobileNetV2 branch enhanced with Squeeze-and-Excitation attention and a custom convolutional branch, with feature-level fusion used to improve discrimination among visually similar disease classes. Under the evaluated mixed-source image-level protocol, the proposed model achieved 97.28% test accuracy, with macro-precision, macro-recall, and macro-F1-score values of 96.75%, 96.69%, and 96.66%, respectively. Calibration and qualitative Grad-CAM analyses were used to examine model behaviour. The image-level partitioning strategy, together with the absence of systematically retained plant- and acquisition-group identifiers, prevented verification of independence at the plant, farm, acquisition-session, state, collection-phase, or source level. Although combining field-acquired and public images increased acquisition diversity within the assembled dataset, it did not constitute source-independent external validation. The reported performance should therefore be interpreted as internal mixed-source image-level performance rather than evidence of independent field-domain generalisation. The model has a compact size of approximately 15.28 MB, supporting its characterisation as architecturally lightweight; however, practical mobile or edge deployment remains to be established through direct device-level evaluation.

Keywords

Tomato leaf disease, MobileNetV2, Attention mechanism, Feature fusion, Deep learning, Precision agriculture

Article History

Received: 09 July 2026
Revised: 17 September 2026
Accepted: 30 September 2026
Available Online: 13 October 2026

Copyright

© 2026 by the authors. This article is published by the Cultech Publishing Sdn. Bhd. under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0

1. Introduction

Tomato (*Solanum lycopersicum*) ranks among the most economically significant vegetable crops cultivated worldwide, valued both for its nutritional composition and its broad role in household food security, commercial agriculture, and industrial food processing [1]. Rich in vitamins C and K, lycopene, and other dietary antioxidants, the crop occupies a central position in the diets of populations across tropical, subtropical, and temperate regions alike [2,3]. In Nigeria, tomato production constitutes a critical pillar of the agricultural economy, sustaining millions of smallholder farmers engaged in both subsistence cultivation and commercial supply chains that feed urban markets [4]. Yet despite its agronomic and economic relevance, tomato productivity in the country continues to be undermined by a wide spectrum of foliar diseases whose effects on yield and fruit quality are well documented and, in many growing seasons, economically devastating [5,6].

Tomato plants are susceptible to numerous pathological conditions arising from fungal, bacterial, viral, and arthropod-related agents. Diseases such as Early blight (*Alternaria solani*), Late blight (*Phytophthora infestans*), Bacterial spot (*Xanthomonas spp.*), Septoria leaf spot, Target spot, and Tomato Yellow Leaf Curl Virus collectively represent the primary causes of yield loss across tomato-growing regions in sub-Saharan Africa [7,8]. Under the warm, humid conditions that prevail in many parts of Nigeria, these pathogens proliferate rapidly and can devastate entire crop cycles within relatively short periods. What makes disease management particularly difficult in these settings is that several of the most destructive conditions produce overlapping visual symptoms, including similar lesion morphologies, comparable patterns of necrosis, and analogous chlorotic responses, which confound accurate identification, especially during the early stages of infection when intervention would be most effective [9-11].

Conventional disease diagnosis in Nigerian farming communities continues to rely predominantly on visual inspection conducted by farmers themselves or, where available, by agricultural extension workers. While this approach is practical in principle, it carries well-known limitations [12,13]. The process is inherently subjective, heavily dependent on individual expertise, and vulnerable to misidentification in cases where disease symptoms are atypical or where multiple conditions co-occur on the same plant. More critically, access to trained plant pathologists remains severely restricted across rural farming landscapes, a constraint that routinely results in delayed diagnosis, inappropriate treatment decisions, and avoidable crop losses. These structural limitations in Nigeria's plant health infrastructure underscore the need for scalable, objective, and accessible alternatives to manual disease assessment [14-16].

The rapid development of deep learning over the past decade has opened a genuinely promising avenue for addressing this diagnostic gap. Convolutional neural networks can extract hierarchical visual representations directly from raw image data [17-20]. This capability reduces reliance on the handcrafted feature engineering associated with earlier machine learning approaches [21-23]. Seminal work by Mohanty et al. [24] established a proof of concept for deep CNN-based plant disease classification at scale using the PlantVillage benchmark dataset, demonstrating that such models could achieve high accuracy across dozens of disease categories spanning multiple crop species. Ferentinos [25] subsequently reinforced this finding across a broader range of architectural configurations, while Too et al. [26] provided a comparative analysis of contemporary deep learning architectures under plant disease classification conditions, reporting particularly strong results from transfer learning strategies. In the years following these foundational studies, pretrained architectures including AlexNet, VGGNet, ResNet, DenseNet, EfficientNet, and MobileNet have all been applied to tomato disease recognition tasks with varying degrees of success.

Notwithstanding this body of progress, a careful review of the existing literature reveals several persistent limitations that constrain the practical relevance of many published systems. Perhaps the most consequential is the heavy reliance on controlled, laboratory-curated datasets, chief among them PlantVillage, which are characterised by uniform illumination, plain backgrounds, and images captured under standardised conditions that bear little resemblance to the photographic variability encountered in real agricultural environments [27-30]. Models trained exclusively on such datasets frequently exhibit pronounced performance degradation when evaluated on field-collected images affected by inconsistent lighting, partial occlusion, complex vegetative backgrounds, variable image resolutions, and the full range of disease severity encountered in actual growing conditions [31-33]. The gap between laboratory performance and field deployability is not merely a technical inconvenience; it directly limits the translational value of these systems for farmers and agricultural practitioners in settings such as Nigeria.

A second limitation concerns computational scalability. Many high-performing deep learning architectures demand considerable memory resources and processing capacity that simply cannot be met by the mobile and edge computing environments in which practical agricultural Artificial Intelligence (AI) systems would most plausibly operate [34,35]. Although large-scale models such as VGGNet and deep ResNet variants achieve impressive classification accuracies under controlled conditions, their deployment requirements render them impractical for use on low-cost smartphones or field-based devices without substantial infrastructure support. This consideration has motivated growing interest in lightweight architectures, among which MobileNetV2 has attracted considerable attention [36,37]. Its depthwise separable convolution design substantially reduces the number of trainable parameters relative to conventional convolutions, enabling competitive classification performance within stringent computational budgets.

A third and related limitation is the tendency of conventional transfer learning approaches to treat all feature channels with equal importance, irrespective of their discriminative relevance to the target classification task. For tomato diseases

with pronounced visual similarity, Early blight and Late blight being perhaps the most frequently cited example, the capacity to selectively emphasise diagnostically informative regions within learned feature representations can meaningfully improve classification outcomes [38-40]. Attention mechanisms offer a principled framework for achieving this selective emphasis. The Squeeze-and-Excitation (SE) network proposed by Hu et al. [41] introduced channel-wise feature recalibration as a lightweight, pluggable module that could be integrated into standard convolutional architectures with minimal overhead, yielding consistent improvements in representational quality. Subsequent studies applying attention-enhanced architectures to agricultural image classification tasks have reported benefits from this approach, although the extent to which such mechanisms improve performance under genuinely heterogeneous imaging conditions remains an open empirical question.

Recent lightweight attention-based plant disease classification models have pursued several routes to balance predictive performance with computational efficiency. XSE-TomatoNet, an explainable tomato leaf disease classification model based on EfficientNetB0, incorporates SE blocks and multi-scale feature fusion and reported an accuracy of 99.11% on a publicly available tomato leaf dataset [28]. Sun et al. [42] combined inverted residual modules with Efficient Channel Attention (ECA) to develop an ultra-lightweight tomato disease recognition model for complex environments, achieving 97.84% accuracy with 0.418 million parameters. Batool et al. [43] introduced Lightweight Depthwise Separable Convolution with Spatial Attention (LWDSC-SA), which integrates depthwise separable convolutions and spatial attention within a compact architecture and achieved 98.70% accuracy with 2.4 million parameters on the 38-class PlantVillage dataset. The Dual-Attention Multi-Scale Lightweight Network (DAMSLNet) proposed by Deng et al. [44] combines depthwise separable convolutions, multi-scale feature extraction, deep feature fusion, and dual-attention mechanisms and demonstrated competitive performance across four plant disease datasets. More recently, Adithiyaa et al. [45] proposed the Attentive and Lightweight Network (ALNet), which integrates Fire modules, residual learning, SE attention, and spatial attention within a 0.17-million-parameter model evaluated on grapevine, apple, and cherry leaf datasets. Although these studies demonstrate the value of lightweight attention mechanisms, their reported accuracies should not be interpreted as directly comparable rankings because they differ in crop species, number of disease classes, image acquisition conditions, dataset composition, and experimental protocols.

MultiBranchFusionNet is positioned within this body of work as a particular integration of established lightweight and attention-based components rather than as the introduction of a new convolutional operator, attention mechanism, or fusion principle. The framework uses two complementary and asymmetric feature-learning pathways. The first employs an ImageNet-pretrained MobileNetV2 backbone followed by SE attention for channel-wise feature recalibration, while the second uses a purpose-built convolutional neural network to learn task-specific visual representations directly from the tomato leaf dataset. The resulting representations are concatenated before classification so that transferable high-level features and task-specific convolutional features contribute jointly to the prediction. The contribution therefore lies in the deliberate engineering organisation of these established components into the proposed asymmetric dual-branch framework and in its internal evaluation using a heterogeneous dataset that combines controlled public images with 9,312 field-acquired images collected under variable conditions in South-West Nigeria.

The principal contributions of this study are as follows:

(1) A heterogeneous tomato leaf disease dataset comprising 20,245 images distributed across ten classes was assembled for model development and evaluation. The dataset combines 10,933 publicly available images with 9,312 field-acquired images collected from farms in Ogun, Oyo, and Ondo States, Nigeria, thereby incorporating variations in illumination, background, leaf orientation, symptom severity, and image-capture conditions that are often underrepresented in controlled benchmark datasets. Here, “heterogeneous” denotes diversity in image provenance and acquisition conditions within the assembled dataset and does not imply that cross-source or external-domain generalization was independently validated.

(2) A lightweight asymmetric dual-branch architecture, termed MultiBranchFusionNet, was developed to integrate complementary feature representations. An ImageNet-pretrained MobileNetV2 branch enhanced with SE attention captures transferable high-level features, while a separate custom CNN branch learns task-specific visual representations directly from the tomato leaf dataset. Their concatenation before classification provides the implementation-level distinction of the proposed framework, in which transferable pretrained representations and task-specific convolutional features are combined within a single asymmetric dual-branch architecture.

(3) The architectural components were examined through an exploratory single-run ablation analysis involving a MobileNetV2 configuration, an attention-enhanced variant, a dual-branch variant without SE attention, and the complete MultiBranchFusionNet. The analysis is used descriptively to document the architecture-development process rather than as a repeated-run statistical attribution of component effects. Development-set-only five-fold cross-validation was conducted using matched folds for MultiBranchFusionNet and MobileNetV2, while the fixed test subset was excluded from all cross-validation operations. Per-class evaluation, confusion-matrix analysis, bootstrap confidence intervals, and fold-wise statistical comparison were employed to provide complementary evidence concerning performance and stability within the evaluated dataset.

(4) The computational characteristics of the proposed framework were quantified using total, trainable, and non-trainable parameter counts, model size, floating-point operations, and training time. With 2.96 million parameters, a model size of

15.28 MB, and a held-out test accuracy of 97.28%, MultiBranchFusionNet exhibits a compact architecture-level profile that motivates further deployment-oriented evaluation, while reproducible device-level inference performance remains to be established.

The remainder of this paper is structured as follows. Section 2 describes the materials and methods adopted in the study, encompassing dataset assembly and preprocessing, comparative model evaluation, and the design and training of the proposed MultiBranchFusionNet architecture. Section 3 presents the experimental results alongside a detailed performance analysis and discusses the principal findings and their implications for practical agricultural AI deployment. Section 4 concludes the study and identifies directions for future research.

2. Materials and Methods

2.1 Dataset Acquisition and Description

The image dataset assembled for this study was derived from two complementary sources: a field-acquired collection obtained from tomato farms in South-West Nigeria and a publicly available tomato leaf dataset. The field collection was conducted in two phases. The first phase took place between July and September 2024, while the second was carried out between February and August 2025. In total, images were collected from 26 farms distributed across three states: 8 farms in Ogun State, 12 in Oyo State, and 6 in Ondo State. Before image acquisition commenced, access to each farm and permission to photograph tomato plants for research purposes were obtained from the farm manager or other responsible personnel present at the location. The field-acquisition activity involved photographing tomato plants and leaves and did not involve the recruitment of human participants or the collection of personal data from farm personnel.

A purposive, symptom-targeted sampling strategy was adopted during the field survey. Plants and leaves were deliberately selected when they exhibited visible symptoms considered consistent with the disease categories investigated in the study. Images were captured using an iPhone 14 Pro Max equipped with a 48-megapixel main camera, a 24 mm equivalent focal length, an f/1.78 aperture, second-generation sensor-shift optical image stabilization, and 100% Focus Pixels. The leaves remained attached to the plants during image acquisition and were photographed on-site from multiple viewing angles and at varying distances. No artificial background, fixed illumination system, or standardised camera orientation was imposed. Consequently, the field images retained naturally occurring variations in illumination, shadow distribution, background vegetation, leaf orientation, symptom severity, viewing distance, and disease presentation. The field-acquired subset comprised 9,312 images. Multiple images were sometimes captured from the same plant, using different viewing angles and distances to represent variation in symptom appearance and field imaging conditions. The number of individual plants represented in the field collection was not systematically recorded during acquisition. The original photographs also had varying image resolutions because they were acquired under natural field conditions rather than with a fixed image-resolution protocol. All images were subsequently resized to 224 × 224 pixels during preprocessing to provide a consistent input size for the neural network.

The second source was the PlantVillage Tomato Leaf Dataset made publicly accessible through Kaggle [46]. Following curation and class harmonisation, 10,933 images from this repository were retained to provide additional examples of the relevant tomato leaf conditions and to supplement classes that were less extensively represented in the field collection. The public images were accessed through the cited Kaggle repository and used in accordance with the access and use conditions presented by the repository. Because the specific repository listing does not identify a named reuse licence, no additional redistribution or licensing rights beyond those provided by the source repository are asserted in this study.

Field-image labels were assigned through a two-stage, symptom-based annotation procedure. During image acquisition, field enumerators assigned preliminary disease categories based on the visible symptoms observed on the leaves, and these preliminary labels were recorded at the respective source farms. Images presenting uncertain, overlapping, or insufficiently distinctive symptoms were referred to a co-author who is an Associate Professor, holds a PhD, and is a plant pathologist. The co-author provided expert symptom-based review and final class adjudication for cases requiring further assessment. Images that remained non-diagnostic, visually ambiguous, or suggestive of more than one condition after assessment were excluded during dataset curation rather than retained under an uncertain class assignment.

The field annotations were based on visual symptom assessment and were not confirmed through laboratory culture, pathogen isolation, serological testing, molecular analysis, or other laboratory diagnostic procedures. The resulting labels should therefore be interpreted as symptom-based expert assessments rather than laboratory-confirmed disease diagnoses. A separate annotation audit log recording the exact number of images reviewed by the plant-pathology expert, the number of preliminary labels changed during adjudication, and the frequency of initial disagreement was not retained. Consequently, inter-rater agreement and adjudication-frequency statistics cannot be reconstructed retrospectively.

Before dataset partitioning, corrupted files and images considered unsuitable for reliable visual classification were excluded. Following curation, the dataset contained 20,245 tomato leaf images distributed across ten classes: Bacterial Spot, Early Blight, Healthy, Late Blight, Leaf Mold, Grey Leaf Spot, Septoria Leaf Spot, Spider Mites, Target Spot, and Yellow Leaf Curl Virus. Of these, 9,312 images originated from the field collection and 10,933 were obtained from the PlantVillage source.

The complete dataset was divided into training, validation, and test subsets using stratified image-level sampling with a 70:20:10 allocation. The partitioning produced 14,171 training images, 4,049 validation images, and 2,025 test images. Stratification approximately preserved the relative representation of the ten disease categories across the three subsets. The training subset was used for model optimisation, the validation subset for convergence monitoring and checkpoint selection, and the test subset for final internal performance evaluation.

The field subset comprised 9,312 images collected from 26 farms across Oyo, Ogun, and Ondo States during two acquisition phases: July to September 2024 and February to August 2025. The experimental partition was performed at image level and did not prospectively enforce separation by individual plant, farm, acquisition session, state, collection phase, or image source. Unique plant identifiers were not systematically recorded, and multiple photographs were sometimes obtained from the same plant. Consequently, the reported test results represent internal image-level evaluation of the assembled mixed-source dataset rather than independent evaluation across these acquisition groups.

The PlantVillage and field-acquired components also differ substantially in background, illumination, framing, viewing conditions, and image-acquisition characteristics. Because the available evaluation outputs do not preserve a sufficiently auditable prediction-level linkage to image source, field-only and PlantVillage-only performance measures are not reconstructed retrospectively. The state- and collection-phase composition of the field subset is presented in Table 1, while the class-wise train, validation, and test distributions are presented in Table 2. Representative examples from the field-acquired and PlantVillage image collections are shown in Figure 1.

Table 1. State- and collection-phase distribution of the field-acquired image subset.

State	Number of farms	Phase 1: Jul-Sep 2024	Phase 2: Feb-Aug 2025	Total field images
Oyo	12	1,921	2,440	4,361
Ogun	8	1,215	1,507	2,722
Ondo	6	1,044	1,185	2,229
Total	26	4,180	5,132	9,312

The counts in Table 1 describe the provenance of the complete field-acquired subset and do not represent acquisition-group-separated experimental partitions.

Table 2. Class-wise distribution across the dataset partitions.

Disease class	Total	Training	Validation	Test
Yellow Leaf Curl Virus	5,357	3,750	1,071	536
Bacterial Spot	3,166	2,216	633	317
Late Blight	1,881	1,317	376	188
Septoria Leaf Spot	1,771	1,240	354	177
Spider Mites	1,676	1,173	335	168
Healthy	1,650	1,155	330	165
Early Blight	1,622	1,135	325	162
Target Spot	1,404	983	281	140
Leaf Mold	952	666	191	95
Grey Leaf Spot	766	536	153	77
Total	20,245	14,171	4,049	2,025

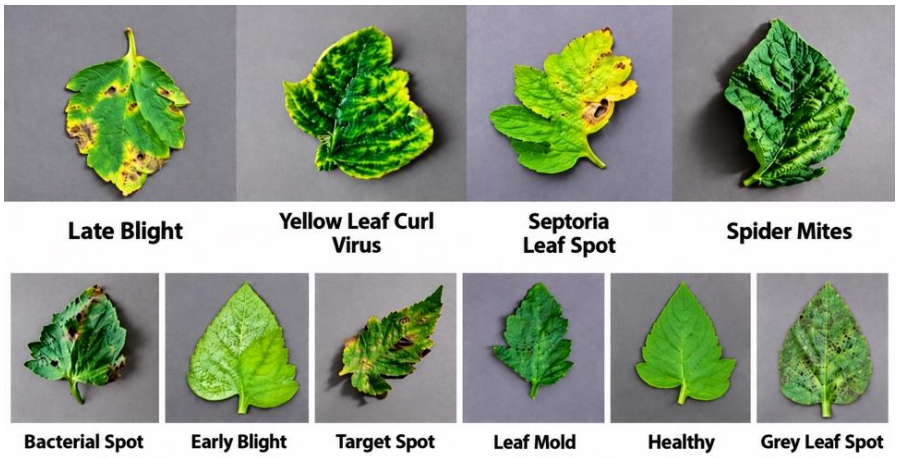


Figure 1. Representative samples from the tomato leaf dataset.

The class distribution shown in Table 2 reveals a meaningful difference in representation across the ten categories. Yellow Leaf Curl Virus contained 5,357 images, whereas Grey Leaf Spot contained 766 images, making the largest class approximately seven times the size of the smallest. This imbalance creates the possibility that aggregate accuracy may be influenced disproportionately by the more frequently represented categories. The recorded final training configuration did not use class weighting, class-balanced sampling, or explicit minority-class oversampling. Consequently, macro-averaged precision, recall, and F1-score are reported alongside class-specific performance and support values so that predictive behaviour in the smaller classes remains visible rather than being obscured by overall accuracy. Online image augmentation was used during training to increase visual diversity, but it was not implemented as a class-specific rebalancing mechanism.

2.2 Image Preprocessing and Augmentation

All images were resized to 224 × 224 pixels before model input. Pixel intensities were then rescaled from the original [0,255] range to [0,1]. This uniform rescaling was the preprocessing implemented in the recorded experiments for both the pretrained and custom feature-extraction pathways. The MobileNetV2-specific preprocess_input transformation, which scales input values to the [-1,1] range used by the standard ImageNet-pretrained implementation, was not applied in the recorded pipeline. Accordingly, all performance results reported in this study correspond to the [0,1] input scaling used during model development and evaluation.

Online image augmentation was applied only to the training data and was implemented through the training-data generator. The transformations comprised random rotations of up to 30°, horizontal flipping, and nearest-neighbour filling of pixels introduced at image boundaries by geometric transformation. The augmentation operations were applied stochastically to images supplied by the training generator rather than selectively to predetermined minority classes. Consequently, no fixed expanded dataset was created, no fixed number of augmented samples existed for individual classes, and mini-batches were not explicitly balanced by class. Validation and test images were not augmented and underwent only resizing and the [0,1] rescaling described above.

Nearest-neighbour filling was the boundary-handling rule used in the implemented augmentation pipeline. This procedure fills newly exposed regions using the value of the nearest valid pixel rather than generating interpolated colour values. Although computationally simple and consistent with the recorded implementation, it can create artificial edge patterns at transformed image boundaries. No separate augmentation ablation was conducted to establish that nearest-neighbour filling was preferable to alternative filling strategies; this possibility is therefore acknowledged as a methodological limitation.

2.3 Preliminary Backbone Evaluation

Before selecting the pretrained pathway for MultiBranchFusionNet, MobileNetV2, ResNet50, MobileNetV3-Small, EfficientNetB0, and MobileNetV3-Large were examined during preliminary model development. The screening provided an engineering basis for backbone selection using the assembled dataset and the recorded classification and model-size measures. Because complete architecture-specific optimisation, fine-tuning, checkpoint, and repeated-run records were not retained uniformly across all candidates, the comparison is treated as a preliminary selection exercise rather than an exhaustively controlled benchmark. The recorded screening results are presented in Table 3.

Table 3. Preliminary performance screening of candidate pretrained backbone architectures.

Model	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)	Size (MB)
EfficientNetB0	88.00	85.00	86.00	98.00	16.71
MobileNetV3-Small	76.00	54.00	64.00	94.00	4.33
MobileNetV3-Large	79.00	62.00	71.00	96.00	7.93
ResNet50	81.00	72.00	76.00	97.00	94.77
MobileNetV2	93.50	92.00	93.00	99.30	9.41

The screening used the common [0,1] image-rescaling pipeline described in Section 2.2. Precision, recall, and F1-score were derived from predicted class labels, while ROC-AUC was calculated from class-probability outputs. MobileNetV2 provided the strongest recorded combination of classification measures and compact model size and was therefore selected as the pretrained pathway. Its recorded held-out accuracy during this preliminary stage was 92.70%. These results document the engineering basis for backbone selection and are not interpreted as evidence of universal superiority over independently optimised alternatives.

Based on the preliminary backbone screening, MobileNetV2 was selected as the pretrained pathway for subsequent architecture development because it combined strong recorded classification performance with a comparatively compact model profile. As a lightweight transfer learning model, MobileNetV2 relies on knowledge acquired from ImageNet pretraining to extract general visual patterns such as edges, textures, and object shapes. These generic features have proven effective across many image classification tasks. However, tomato leaf disease identification is considerably more challenging because many diseases produce symptoms with similar lesion shapes, colour changes, and spatial

distributions. These subtle visual similarities, together with the variability introduced by real field conditions, make accurate discrimination more difficult. Differences in lighting, background clutter, leaf orientation, occlusion, and disease severity further increase the complexity of the classification task. Consequently, representations learned solely from ImageNet may not always capture the fine-grained disease characteristics needed for reliable recognition.

To overcome these challenges, this study introduces a dual-branch feature extraction framework in which two complementary networks learn different aspects of the input images simultaneously. The first branch employs the pretrained MobileNetV2 model to retain the advantages of transfer learning by extracting robust high-level semantic features while retaining a comparatively compact computational profile that motivates subsequent deployment-oriented evaluation. Alongside this branch, a custom convolutional network is trained directly on the tomato leaf dataset to learn features that are specific to the disease classification task. Unlike the pretrained backbone, this branch is not influenced by biases inherited from large-scale natural image datasets, allowing it to focus on characteristics that are unique to tomato leaf diseases. By combining these two learning pathways, the model is able to capture complementary information that would be difficult for either branch to learn independently.

To further improve feature learning, a SE attention mechanism was incorporated into the pretrained MobileNetV2 branch. The module adaptively recalibrates the feature channels generated by the backbone, assigning greater importance to channels carrying disease-relevant information while reducing the influence of less informative responses. The custom convolutional branch does not contain an SE block and instead learns task-specific representations directly from the tomato leaf images.

The features extracted from both branches are then combined through a feature fusion process, enabling the classifier to benefit from the strengths of each pathway. Rather than depending on a single feature representation, the fused feature space integrates the broad semantic knowledge provided by transfer learning with the task-specific information learned directly from the tomato dataset. This richer representation improves the model's ability to distinguish between visually similar disease categories. The recorded performance of selected architectural variants is examined descriptively through the exploratory ablation analysis presented in Section 3.6.

2.4 MultiBranchFusionNet Architecture

Based on the motivation discussed in the previous section, MultiBranchFusionNet was designed to address the limitations identified in conventional single-backbone transfer learning models while retaining a compact computational profile intended for deployment-oriented evaluation. The proposed architecture consists of two parallel feature extraction branches that learn complementary representations from the input images. The first branch uses a pretrained MobileNetV2 backbone to exploit the rich visual knowledge gained from large-scale ImageNet pretraining, providing robust high-level feature representations. The second branch is a custom convolutional network trained directly on the tomato leaf dataset to capture disease-specific characteristics that may not be adequately represented by generic pretrained features.

The information learned by both branches is subsequently combined through a feature fusion mechanism, while a channel-wise attention module is incorporated into the MobileNetV2 branch to enhance the most informative feature channels and suppress less relevant ones. By integrating these complementary learning strategies, the proposed architecture aims to improve the recognition of visually similar tomato leaf diseases without significantly increasing computational complexity. The overall architecture of MultiBranchFusionNet is presented in Figure 2.

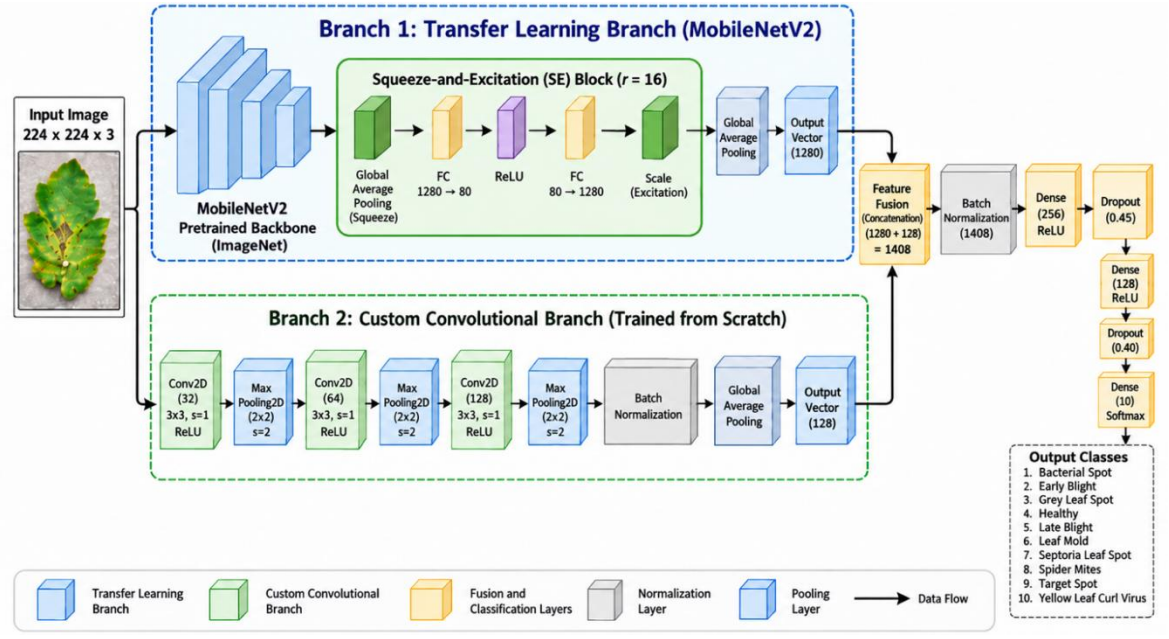


Figure 2. Architecture of the proposed MultiBranchFusionNet.

2.4.1 MobileNetV2 Feature Extraction Branch

The first branch of MultiBranchFusionNet employs the pretrained MobileNetV2 architecture as a feature extractor. The pretrained backbone was kept frozen during training, with learning restricted to the task-specific components of the proposed network. The trainable components comprised the SE block, the three convolutional layers and batch-normalisation layer of the custom branch, the fusion batch-normalisation layer, and the fully connected classification layers, while the global average pooling and dropout operations contained no trainable parameters.

The input supplied to this branch followed the common [0,1] rescaling described in Section 2.2; no additional MobileNetV2-specific preprocess_input transformation was applied before feature extraction. Thus, the frozen ImageNet-pretrained backbone operated on the same rescaled image representation supplied to the parallel custom convolutional branch.

MobileNetV2 is built around depthwise separable convolutions and inverted residual blocks with linear bottlenecks, a design that reduces computational requirements while preserving useful feature representations across multiple spatial scales. In the context of tomato disease classification, this branch extracts high-level visual representations associated with leaf surface texture, lesion distribution, colour variation, and broader structural characteristics learned from ImageNet pretraining.

MobileNetV2 was configured with an input size of 224 × 224 × 3, α = 1.0, and include_top = False. For the specified input resolution, the backbone produces a 7 × 7 × 1280 feature map. A single SE block is applied to this feature map before global average pooling. The SE block uses a reduction ratio of 16 and therefore reduces the 1,280-channel representation to 80 units before restoring it to 1,280 units. ReLU activation is used in the reduction layer, while sigmoid activation is used in the expansion layer. The recalibrated feature map is then passed through global average pooling to produce a 1,280-dimensional feature vector. The detailed configuration of this branch and the custom convolutional branch is presented in Table 4.

Table 4. Layer configuration and feature dimensions of the MultiBranchFusionNet feature-extraction branches.

Branch	Layer/Operation	Configuration	Output dimension
Input	Input image	RGB image	224 × 224 × 3
MobileNetV2	MobileNetV2	α = 1.0, ImageNet pretrained, include_top = False, frozen	7 × 7 × 1280
MobileNetV2	SE block	One block, reduction ratio = 16	7 × 7 × 1280
MobileNetV2	Global Average Pooling	Global spatial averaging	1280
Custom CNN	Conv2D	32 filters, 3 × 3 kernel, stride 1, same padding, ReLU	224 × 224 × 32
Custom CNN	MaxPooling2D	2 × 2 pool, stride 2	112 × 112 × 32
Custom CNN	Conv2D	64 filters, 3 × 3 kernel, stride 1, same padding, ReLU	112 × 112 × 64
Custom CNN	MaxPooling2D	2 × 2 pool, stride 2	56 × 56 × 64
Custom CNN	Conv2D	128 filters, 3 × 3 kernel, stride 1, same padding, ReLU	56 × 56 × 128
Custom CNN	MaxPooling2D	2 × 2 pool, stride 2	28 × 28 × 128
Custom CNN	Batch Normalisation	Applied after third pooling operation	28 × 28 × 128
Custom CNN	Global Average Pooling	Global spatial averaging	128

The SE mechanism performs channel-wise feature recalibration through squeeze and excitation operations. In the squeeze stage, global average pooling compresses the spatial dimensions of each feature map into a single scalar descriptor for each channel. For channel c, this operation is defined in Eq. (1), where $x_c(i, j)$ represents the activation at spatial position (i, j) in channel c, and H and W denote the height and width of the feature map respectively.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c\ (i, j) \tag{1}$$

In the excitation stage, the resulting channel descriptor vector is passed through a two-layer fully connected sub-network to generate channel-wise scaling weights. For the implemented architecture, the 1,280-dimensional descriptor is reduced to 80 units and then expanded back to 1,280 units. The excitation operation is expressed in Eq. (2), where W_1 and W_2 are learnable weight matrices, δ denotes the ReLU activation function, and σ denotes the sigmoid activation function.

$$s = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

The recalibrated feature representation is subsequently obtained by scaling the original feature map channel-wise using the computed weights. This channel-wise multiplication operation is defined in Eq. (3), where sc denotes the scalar weight assigned to channel c and xc denotes the corresponding unscaled feature map.

$$\tilde{x}_c = s_c \cdot x_c \tag{3}$$

Collectively, the operations described in Eqs. (1)-(3) allow the model to assign greater emphasis to informative feature channels while reducing the influence of less informative channels.

2.4.2 Custom Convolutional Feature Extraction Branch

The second branch was constructed from scratch as a custom convolutional network consisting of convolutional, max-pooling, batch-normalisation, and global average pooling operations. Its primary purpose was to learn task-specific disease representations directly from the tomato dataset, without the inductive biases introduced by ImageNet pretraining.

The custom convolutional branch consists of three convolutional layers with 32, 64, and 128 filters, respectively. Each convolutional layer uses a 3 × 3 kernel, a stride of 1, same padding, and ReLU activation, followed by a 2 × 2 max-pooling operation with a stride of 2. The first convolutional layer therefore produces a 224 × 224 × 32 feature map, which is reduced to 112 × 112 × 32 after max pooling. The second convolutional layer produces a 112 × 112 × 64 feature map, followed by pooling to obtain 56 × 56 × 64. The third convolutional layer produces a 56 × 56 × 128 feature map, followed by a third pooling operation that reduces the spatial dimension to 28 × 28 × 128. Batch normalisation is then applied before global average pooling, which converts the resulting feature maps into a 128-dimensional task-specific feature vector. No SE block, flattening operation, dropout layer, or fully connected layer is used within the custom convolutional branch. The complete layer configuration and output dimensions of both feature-extraction branches are presented in Table 4.

2.4.3 Feature Fusion Strategy

The feature vectors produced independently by the SE-enhanced MobileNetV2 branch and the custom convolutional branch were combined through concatenation to form a unified fused representation. Let F_{m} denote the output feature vector of the MobileNetV2 branch and F_{c} the corresponding output of the custom branch. The construction of the fused representation F_f through channel-wise concatenation is formally defined in Eq. (4), where [;] denotes vector concatenation along the feature dimension.

$$F_f = [F_m;\ F_c] \tag{4}$$

The fused feature vector produced by Eq. (4) has a dimension of 1,408, resulting from the concatenation of the 1,280-dimensional MobileNetV2 feature vector and the 128-dimensional custom convolutional feature vector. Batch normalisation is then applied to the fused representation, followed by a fully connected layer with 256 units and ReLU activation, dropout with a rate of 0.45, a second fully connected layer with 128 units and ReLU activation, and dropout with a rate of 0.40. The final classification layer contains ten units and uses Softmax activation to generate the probability distribution over the ten disease categories. The feature dimensions and classification-layer configuration are presented in Table 5.

Table 5. Feature fusion and classification configuration of MultiBranchFusionNet.

Stage	Layer/Operation	Configuration	Output dimension
MobileNetV2 branch	Feature vector	SE-refined global average pooled representation	1280
Custom CNN branch	Feature vector	Global average pooled representation	128
Fusion	Concatenation	1280 + 128	1408
Fusion	Batch Normalisation	Applied to fused representation	1408
Classifier	Dense	256 units, ReLU	256
Classifier	Dropout	Rate = 0.45	256
Classifier	Dense	128 units, ReLU	128
Classifier	Dropout	Rate = 0.40	128
Output	Dense	10 units, Softmax	10

2.5 Training Configuration and Experimental Reproducibility

MultiBranchFusionNet was implemented in Python 3.12 using TensorFlow 2.18.0 and its integrated Keras application programming interface. The supporting software environment comprised NumPy 2.1.0, pandas 2.2.3, Matplotlib 3.11.1, Seaborn 0.13.2, scikit-learn 1.9.0, path 17.1.1, and typing_extensions 4.16.0. These package versions were used for numerical processing, data organisation, model evaluation, visualisation, file-path handling, and supporting Python functionality, as applicable. All model-development and evaluation experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4060 GPU, an Intel Core i7 processor, and 32 GB of RAM. Mixed-bfloat16 precision and Accelerated Linear Algebra compilation were enabled to improve computational efficiency during training.

Model optimisation employed the Adam optimiser with an initial learning rate of 0.0001 and categorical cross-entropy as the loss function. The final training configuration did not use class weighting, class-balanced sampling, explicit minority-class oversampling, SMOTE, or focal loss. Online augmentation was implemented through the training generator to increase visual variability in the training images, but it was not used as a class-specific resampling strategy. Consequently, no class-weight computation formula, class-specific weight values, fixed augmented-image totals, or class-specific augmentation frequencies are applicable to the reported final configuration.

Training was configured for a maximum of 100 epochs. An EarlyStopping callback with a patience of five epochs was included; however, the retained implementation shows that it monitored val_accuracy, whereas the compiled validation metric was logged as val_categorical_accuracy. Consequently, the intended early-stopping criterion did not operate on the available validation categorical-accuracy metric, and training continued to the configured 100-epoch limit. ModelCheckpoint independently monitored val_categorical_accuracy and correctly retained the best-performing checkpoint at epoch 20, where validation categorical accuracy reached 0.96888. A ReduceLROnPlateau callback separately monitored validation loss with a patience of three epochs, a reduction factor of 0.5, and a minimum learning rate of 1×10^{-6}. The saved epoch-20 checkpoint was subsequently loaded for final test evaluation.

The augmentation procedure described above corresponds to the recorded final MultiBranchFusionNet training configuration. Complete architecture-specific and fold-wise augmentation records were not retained consistently for every preliminary experimental comparison; equivalence of augmentation and class-imbalance handling is therefore not assumed for configurations for which it cannot be verified. Table 6 summarises the model's parameter count, storage size, computational complexity, and total training time.

Table 6. Computational characteristics of MultiBranchFusionNet.

Metric	Proposed MultiBranchFusionNet
Total parameters	2.96 M (2,958,426)
Trainable parameters	0.70 M (697,370)
Non-trainable parameters	2.26 M (2,261,056)
Model size	15.28 MB
Computational complexity	1.63 GFLOPs
Total training time	20,378.8 s (≈5.66 h)

MultiBranchFusionNet contained 2,958,426 parameters, comprising 697,370 trainable and 2,261,056 non-trainable parameters, with the latter arising primarily from the frozen MobileNetV2 backbone. The final model occupied 15.28 MB and had a nominal computational complexity of 1.63 GFLOPs per forward pass. A reproducible inference-latency benchmark is not reported because the experimental timing record does not preserve the warm-up procedure, number of measured runs, device-synchronisation procedure, or run-level latency observations required for verification. Consequently, mean latency, throughput, and latency-distribution statistics are not reconstructed retrospectively.

The reported 1.63 GFLOPs represents a nominal arithmetic-operation count and was not adjusted according to numerical precision. Although the mixed_bfloat16 policy may influence memory use and hardware execution efficiency, no matched FP32 benchmark was conducted and no precision-specific performance benefit is claimed. Peak inference memory, energy consumption, device-level latency, and model-conversion or quantisation performance were not measured.

For reproducibility, the main hyperparameters and experimental settings of the final MultiBranchFusionNet configuration are listed in Table 7. These settings describe the recorded final proposed-model run and should not be assumed to apply identically to preliminary backbone screening or other experimental configurations for which complete architecture-specific records were not retained.

Table 7. Hyperparameters and experimental settings adopted for MultiBranchFusionNet.

Parameter	Value
Backbone network	MobileNetV2 (ImageNet pretrained)
Input image size	224 × 224 × 3
Batch size	32
Optimiser	Adam
Initial learning rate	0.0001
Loss function	Categorical cross-entropy
Class-imbalance handling	No class weighting, class-balanced sampling, or explicit minority-class oversampling in the final recorded run; macro and class-specific metrics reported
Output activation	Softmax
Number of output classes	10
Random seed	3000
Mixed precision policy	mixed_bfloat16
XLA acceleration	Enabled
Early stopping	Configured to monitor val_accuracy; patience = 5; did not activate because the compiled metric was val_categorical_accuracy
Learning-rate scheduler	ReduceLROnPlateau; validation-loss monitoring; patience = 3; reduction factor = 0.5; minimum learning rate = 1×10^{-6}
Model checkpoint	Epoch 20; highest validation categorical accuracy = 0.96888; saved checkpoint used for final test evaluation
Data partitioning	Stratified, image-level 70:20:10 split; split random state = 42
Evaluation scope	Internal mixed-source image-level evaluation
Dataset preprocessing	Resize to 224 × 224 pixels; uniform pixel rescaling to [0,1]; MobileNetV2-specific preprocess_input not applied
Data augmentation	Online training-only augmentation through the training generator: random rotations up to 30°, horizontal flipping, and nearest-neighbour filling; no fixed class-specific augmented counts or class-balanced mini-batches
Aggregate evaluation	Macro-precision, macro-recall, and macro-F1-score

The experimental settings reported in Table 7 correspond to the recorded final MultiBranchFusionNet configuration. The stratified image-level partition used a split random state of 42. Separately, a random seed of 3000 was assigned to the Python, NumPy, and TensorFlow random-number generators for the recorded model-training configuration. Augmentation was applied exclusively to the training subset, whereas the validation and test images underwent only resizing and [0,1] rescaling. The validation subset was used for convergence monitoring, learning-rate adjustment, and checkpoint selection during the final MultiBranchFusionNet training run. The test subset was not used for model fitting, validation monitoring, learning-rate scheduling, or checkpoint selection. After training, the selected checkpoint was evaluated on the held-out test subset to obtain the classification report, confusion matrix, ROC curves, and confidence distribution. However, because preliminary model-development stages included evaluation of candidate configurations on the held-out partition, the complete development history does not support a strict claim that the test subset was accessed only once after all architectural and hyperparameter choices had been finalised.

2.6 Performance Evaluation Metrics

Model performance was assessed using a comprehensive set of classification metrics calculated on the held-out test subset. Because the number of images differed considerably across the ten categories, macro-precision, macro-recall, and macro-F1-score were treated as the principal aggregate measures alongside overall accuracy. Macro averaging calculates each metric independently for every class before taking the unweighted mean, thereby assigning equal importance to all categories, irrespective of their support. This limits the extent to which the larger classes can dominate the aggregate results. Class-specific precision, recall, F1-score, and support were also reported to make performance on the underrepresented categories directly visible, while weighted-average results were retained as supplementary descriptive measures.

Area under the receiver operating characteristic curve was computed for each class under a one-versus-rest formulation. Confusion matrix analysis was conducted to identify systematic misclassification patterns, and training and validation convergence histories were examined to assess learning stability across epochs. In all metric definitions that follow, TP,

TN, FP, and FN denote true positive, true negative, false positive, and false negative counts respectively, computed on a per-class basis for multi-class evaluation.

Classification accuracy, which measures the proportion of correctly classified samples relative to the total number of evaluated instances, is formally defined in Eq. (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision, which quantifies the fraction of positive predictions that are genuinely positive and thus reflects a model's tendency to avoid false alarms, is defined in Eq. (6).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Recall, which captures the proportion of actual positive instances that the model successfully identifies and is therefore sensitive to missed detections, is defined in Eq. (7).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

The F1-score, which represents the harmonic mean of precision and recall and provides a single balanced measure of classification performance particularly suited to imbalanced class distributions, is defined in Eq. (8).

$$\text{F1} = 2 \times \frac{\text{Precesion} \times \text{Recall}}{\text{Precesion} + \text{Recall}} \tag{8}$$

Prediction calibration was evaluated using top-label Expected Calibration Error (ECE), categorical negative log-likelihood (NLL), and the multiclass Brier score. For each evaluated image, the maximum softmax probability was treated as the predictive confidence for the ECE calculation, while correctness was determined by agreement between the predicted and ground-truth classes. The confidence range from 0 to 1 was divided into 15 equal-width bins, and ECE was calculated as the sample-weighted absolute difference between mean confidence and empirical accuracy across the non-empty bins. Lower ECE indicates closer agreement between predictive confidence and observed correctness [47].

Categorical NLL was calculated as the sample mean of the negative logarithm of the probability assigned to the ground-truth class, thereby penalising predictions that assign low probability to the correct class, particularly when an incorrect prediction is made with high confidence. The multiclass Brier score was calculated as the sample mean of the sum of squared differences between the predicted probability vector and the corresponding one-hot encoded target vector across the ten classes. Lower values of NLL and Brier score indicate better probabilistic prediction quality. These measures were evaluated alongside ECE because, unlike ECE, they do not depend on the selection of confidence-bin boundaries.

Temperature scaling was applied as a post-hoc calibration procedure using a single positive scalar temperature parameter. The temperature was fitted exclusively using predictions from the validation subset by minimising categorical NLL. The held-out test subset was excluded completely from temperature estimation and was evaluated only after the scalar temperature had been fixed. The validation-fitted temperature was subsequently applied to the test logits before softmax transformation. Because scaling all logits by the reciprocal of the same positive temperature preserves their rank ordering, the procedure modifies predictive confidence without changing the predicted class labels.

To quantify uncertainty surrounding the held-out test results, non-parametric bootstrapping was applied to both the classification and calibration measures. A total of 10,000 bootstrap samples were drawn with replacement from the 2,025 held-out test observations using a fixed random seed of 3000. For the classification analysis, the paired ground-truth and predicted class labels were resampled and accuracy, macro-precision, macro-recall, and macro-F1-score were recalculated for each bootstrap sample. For the calibration analysis, the paired ground-truth labels and corresponding predicted probability vectors were resampled, and ECE, categorical NLL, and multiclass Brier score were recalculated before and after temperature scaling. For each measure, the 2.5th and 97.5th percentiles of the resulting bootstrap distribution were taken as the lower and upper limits of the corresponding 95% confidence interval.

A five-fold cross-validation analysis was also conducted on the development portion of the dataset comprising the original training and validation subsets (n = 18,220), while the fixed test subset of 2,025 images remained excluded from all cross-validation operations. MultiBranchFusionNet and the MobileNetV2 comparison model were evaluated using matched outer folds so that their recorded accuracies could be compared fold by fold. The retained experimental record preserves the fold-wise accuracy values but does not contain a complete per-image fold-assignment file, fold-wise class inventory, or auditable linkage between the folds and plant, farm, acquisition-session, state, source, or pHash-group identifiers. The cross-validation results are therefore treated as a descriptive development-set stability summary rather than as independently reproducible group-aware validation.

For the matched fold-wise comparison, the accuracy difference in each fold was calculated as the MultiBranchFusionNet accuracy minus the corresponding MobileNetV2 accuracy. The mean paired difference and sample standard deviation were calculated directly from the five recorded differences. The 95% CI for the mean paired difference was calculated

from the five fold-wise values using the Student's t distribution. Because the cross-validation training sets overlap, this CI is interpreted cautiously and should not be regarded as an interval derived from five independent experimental replications.

A paired Student's t-test and an exact two-sided Wilcoxon signed-rank test were retained only as supplementary sensitivity analyses. The fold-wise observations are not statistically independent because their training sets overlap, and only five paired values are available. Consequently, neither test is used as definitive evidence of model superiority. For the Wilcoxon analysis, no paired difference was zero. The two equal differences of 3.35 percentage points were assigned tied midranks, and the exact two-sided significance level was obtained from the possible sign assignments of the five non-zero paired differences. Effect interpretation is therefore based primarily on the observed fold-wise differences, their mean, and the associated 95% CI rather than on null-hypothesis significance testing.

To complement the numerical evaluation and examine the spatial evidence contributing to individual predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to representative held-out test images. This analysis was conducted because the SE mechanism used in MultiBranchFusionNet performs channel-wise feature recalibration and does not itself provide spatial attribution of the image regions contributing to a prediction. Grad-CAM was therefore used as a post-hoc visual explanation method rather than as an additional attention component of the classifier.

Because MultiBranchFusionNet contains two parallel convolutional feature-extraction pathways, branch-specific Grad-CAM maps were generated separately for the pretrained and custom branches. For the MobileNetV2 pathway, the terminal spatial activation layer out_relu within the nested mobilenetv2_1.00_224 backbone was used, producing a $7 \times 7 \times 1280$ feature representation. For the custom convolutional pathway, the final convolutional layer conv2d_2**,** which produces a $56 \times 56 \times 128$ representation before the final pooling operation, was used as the target layer.

For each selected image, gradients of the score corresponding to the predicted class were computed with respect to the target convolutional feature maps. The gradients were spatially averaged to obtain channel-importance weights, and a weighted linear combination of the corresponding activation maps was calculated. Rectified linear activation was then applied to retain positive contributions to the predicted class. Each resulting activation map was normalised, resized to the 224×224 input resolution, and superimposed on the original image to enable qualitative comparison between the regions emphasised by the MobileNetV2 and custom-CNN branches.

Representative examples were selected to include both correct and incorrect predictions, with particular attention to Spider Mites, Early Blight, and Target Spot because these categories showed comparatively larger error concentrations in the confusion matrix. Diseased-to-Healthy errors were specifically examined because such errors are practically important in a decision-support setting, where failure to identify visible disease symptoms may delay intervention. Where source identity could be verified from the test-image record, examples from both the field-acquired and PlantVillage components were included. Grad-CAM was interpreted qualitatively as an indication of spatial regions contributing to individual predictions and not as causal evidence that a particular image region alone determined the classifier output.

2.7 Inference Pipeline and Cloud Service Deployment

Following training and evaluation, the model was integrated into an end-to-end inference pipeline for processing submitted tomato leaf images through a remotely hosted prediction service. Each submitted image undergoes format validation, resizing to 224×224 pixels, and pixel-level normalisation before being passed to the trained network. The model returns a probability vector across all ten disease classes, with the final prediction assigned to the class carrying the highest posterior probability value together with an associated confidence score to assist practitioner interpretation. The overall structure of the inference pipeline is illustrated in Figure 3.

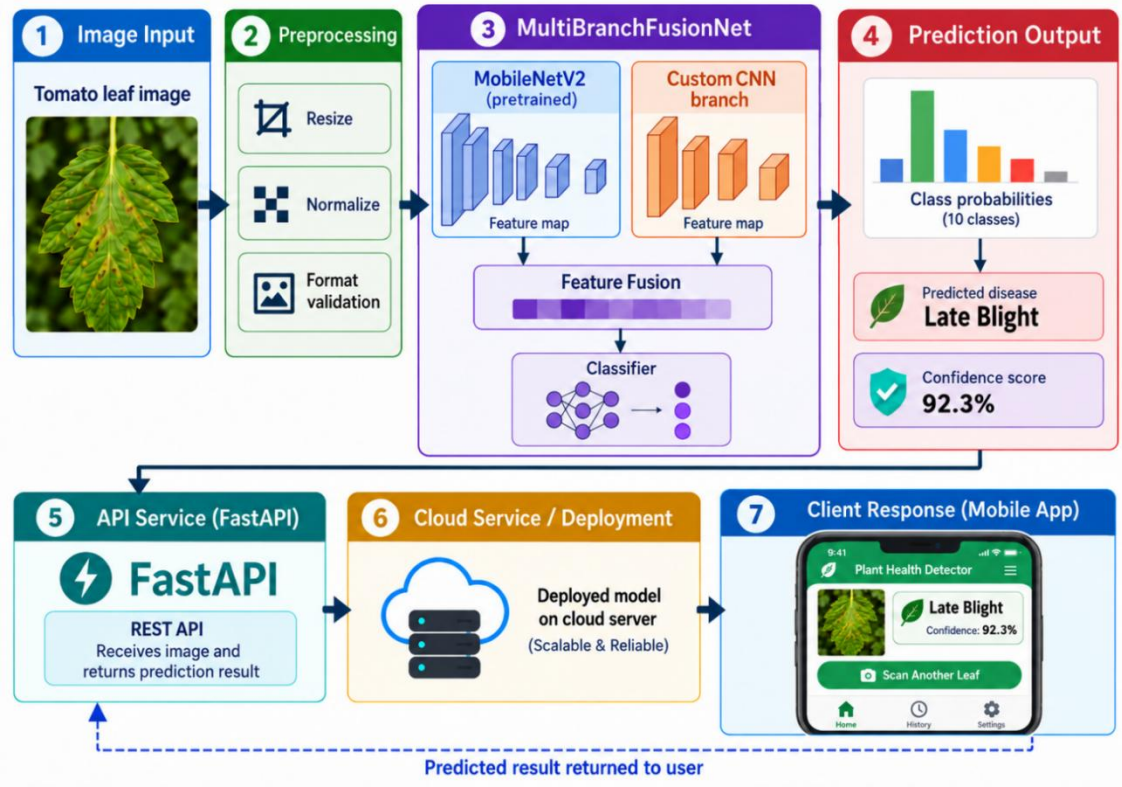


Figure 3. Cloud-based inference-service pipeline for MultiBranchFusionNet.

In this implementation, model inference is executed remotely on the cloud-hosted service, while the user device functions as a client for image submission and result retrieval. The implementation therefore represents cloud-served rather than on-device inference. No deployment-specific model conversion or quantisation, or device-level benchmarking of latency, memory use, energy consumption, or thermal behaviour, was performed. These platform-specific characteristics remain subjects for subsequent deployment evaluation.

3. Results and Discussion

3.1 Classification Performance Analysis

The predictive performance of MultiBranchFusionNet was evaluated across all ten tomato disease categories on the held-out test partition using the saved epoch-20 checkpoint selected by validation categorical accuracy. As presented in Table 8, the model correctly classified 1,970 of the 2,025 test images, corresponding to an accuracy of 97.28% (95% CI: 96.54%-97.98%). The macro-precision was 96.75% (95% CI: 95.87%-97.59%), while the macro-recall and macro-F1-score were 96.69% (95% CI: 95.78%-97.54%) and 96.66% (95% CI: 95.75%-97.50%), respectively. These intervals indicate that the strong overall performance was maintained after accounting for sampling uncertainty within the test set.

Table 8. Classification report for MultiBranchFusionNet.

	Precision	Recall	F1-score	Support
Grey Leaf Spot	1.00	0.99	0.99	77
Bacterial Spot	0.98	0.99	0.98	317
Early Blight	0.97	0.90	0.93	162
Late Blight	0.97	0.99	0.98	188
Leaf Mold	0.97	0.99	0.98	95
Septoria Leaf Spot	0.98	0.99	0.99	177
Spider Mites	0.99	0.88	0.93	168
Target Spot	0.91	0.94	0.93	140
Yellow Leaf Curl Virus	1.00	1.00	1.00	536
Healthy	0.92	1.00	0.96	165
Accuracy			0.97	2025
Macro avg.	0.97	0.97	0.97	2025
Weighted avg.	0.97	0.97	0.97	2025

Because the complete development record does not establish that the test partition remained entirely unexamined throughout all preliminary architecture-selection stages, the reported test performance should be interpreted as internal held-out evaluation rather than as an estimate obtained from a demonstrably single-access post-development test protocol.

These performance measures were calculated for the complete mixed-source test subset and therefore represent aggregate rather than source-specific classification performance. Field-only and PlantVillage-only precision, recall, F1-score, accuracy, and confusion matrices are not reported because the retained prediction outputs cannot be linked retrospectively to source identifiers with sufficient reliability. Accordingly, the aggregate 97.28% accuracy should not be interpreted as evidence that equivalent performance was obtained independently on the field and PlantVillage components.

Given the unequal class distribution, the macro-averaged and class-specific results provide a more informative assessment than accuracy alone. The model achieved a macro-precision of 96.75%, macro-recall of 96.69%, and macro-F1-score of 96.66%, indicating that strong performance was maintained when all ten categories were assigned equal importance. Grey Leaf Spot, the least-represented category, achieved a precision of 1.00, recall of 0.99, and F1-score of 0.99 on 77 test images, while Yellow Leaf Curl Virus achieved 1.00 across the three measures on 536 test images. Nevertheless, estimates derived from smaller class supports are inherently less precise and should be interpreted cautiously. The lower recall of 0.88 for Spider Mites and F1-scores of 0.93 for Early Blight and Target Spot further demonstrate why the aggregate findings should be interpreted alongside the class-specific metrics and support values reported in Table 8.

Overall, the classification results support the effectiveness of the attention-guided dual-branch feature-fusion strategy within the evaluated mixed-source dataset. The 97.28% accuracy should nevertheless be interpreted as internal image-level performance rather than as an estimate of generalisation to previously unseen plants, farms, acquisition sessions, states, collection phases, or image sources. Plant- and acquisition-group independence could not be verified retrospectively; consequently, the magnitude of any optimistic bias arising from correlated images across partitions cannot be quantified from the available records.

3.2 Confusion Matrix Analysis

To further examine prediction behaviour at the class level, the confusion matrix illustrated in Figure 4 was analysed. The matrix provides a detailed breakdown of correctly classified samples alongside the distribution of misclassification cases across all disease categories.

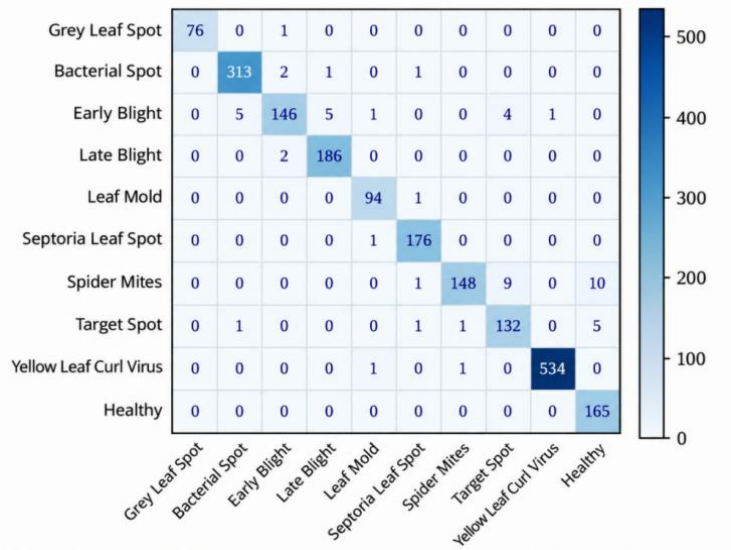


Figure 4. Confusion matrix for MultiBranchFusionNet.

The confusion matrix is dominated by correct classifications, but the remaining errors are concentrated in visually overlapping categories. Early Blight recorded 146 correct predictions from 162 images, with errors distributed across Bacterial Spot, Late Blight, Leaf Mold, Target Spot, and Yellow Leaf Curl Virus. Spider Mites recorded 148 correct predictions from 168 images, including one error to Septoria Leaf Spot, nine to Target Spot, and ten diseased-to-Healthy errors. Target Spot recorded 132 correct predictions from 140 images, including five predictions as Healthy. Diseased-to-Healthy errors are practically important because they may delay intervention. These failure patterns motivated a complementary Grad-CAM analysis to examine whether representative predictions were associated primarily with visible lesion regions, broader leaf structures, or surrounding background features. Particular attention was given to Spider Mites and Target Spot cases predicted as Healthy, as well as misclassifications involving Early Blight and other visually overlapping disease categories. The resulting branch-specific visual explanations are presented and discussed in Section 3.9.1.

3.3 Confidence Calibration Analysis

The probabilistic reliability of MultiBranchFusionNet was evaluated using top-label ECE, categorical NLL, and the multiclass Brier score before and after temperature scaling. In addition to the numerical calibration measures, reliability diagrams were used to examine the relationship between predictive confidence and empirical correctness. The reliability diagrams before and after temperature scaling are presented in Figure 5, providing a direct visual comparison between mean predictive confidence and empirical accuracy across the 15 equal-width confidence bins.

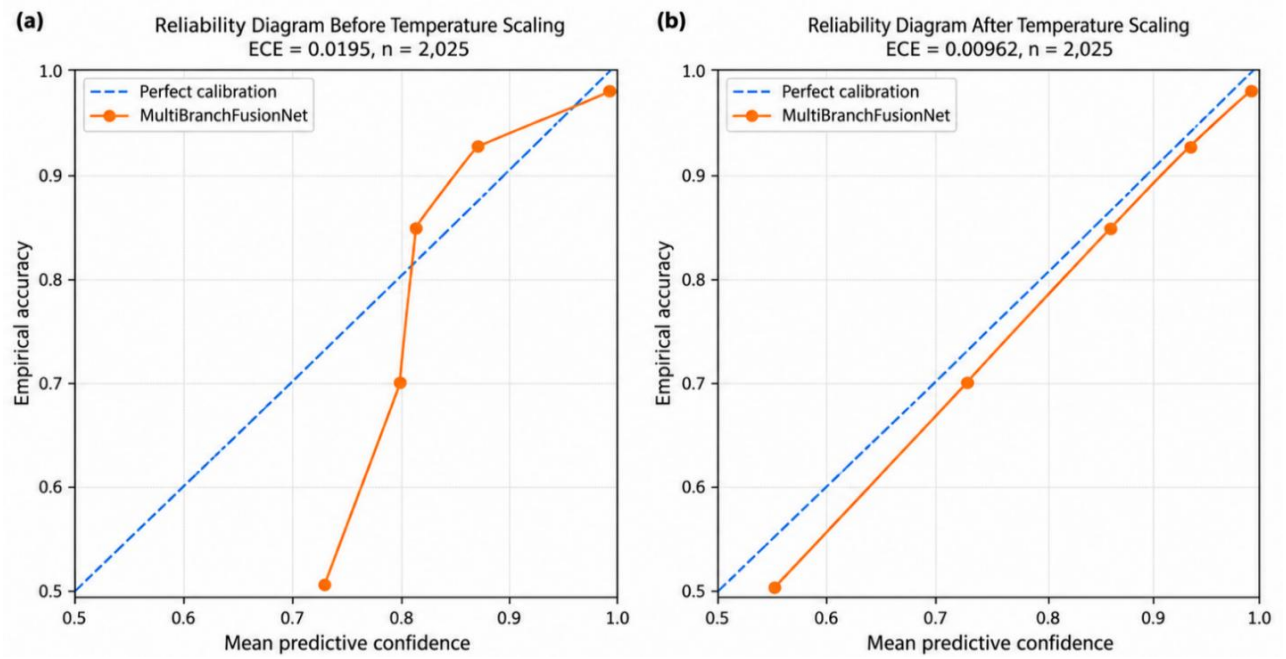


Figure 5. Reliability diagrams for MultiBranchFusionNet on the held-out test subset: (a) before and (b) after temperature scaling.

As shown in Figure 5, the diagonal reference line represents perfect calibration, while the plotted points show empirical accuracy relative to mean predictive confidence for the non-empty bins among the 15 equal-width confidence bins.

The scalar temperature estimated exclusively from the validation predictions was T = 1.18. On the held-out test subset of 2,025 images, top-label ECE decreased from 0.0195 (95% CI: 0.0157-0.0238) before calibration to 0.00962 (95% CI: 0.0067-0.0132) after temperature scaling. This corresponds to an absolute reduction of 0.00988 and a relative reduction of approximately 50.7%.

Categorical NLL decreased from 0.1124 (95% CI: 0.0931-0.1336) before calibration to 0.1047 (95% CI: 0.0872-0.1235) after temperature scaling, corresponding to an absolute reduction of 0.0077. The multiclass Brier score similarly decreased from 0.0492 (95% CI: 0.0406-0.0585) to 0.0460 (95% CI: 0.0380-0.0544), representing an absolute reduction of 0.0032. The reductions in NLL and Brier score provide complementary evidence that the improvement in probabilistic prediction quality was not confined to the bin-dependent ECE measure.

The temperature parameter was fitted only on the validation subset by minimising categorical NLL. The held-out test images were not used during temperature estimation and were evaluated only after the scalar temperature had been fixed. Temperature scaling did not change any predicted test class labels because the transformation preserved the rank ordering of the class logits. Consequently, the model continued to classify 1,970 of the 2,025 test images correctly, and the classification accuracy remained unchanged at 97.28%.

Visual comparison of the reliability diagrams shows that the post-calibration confidence-accuracy relationship lies closer to the ideal diagonal than the corresponding pre-calibration relationship. This pattern is consistent with the reduction in ECE from 0.0195 to 0.00962 and the accompanying decreases in NLL and Brier score. The combined results therefore indicate improved probabilistic agreement within the evaluated internal test distribution.

A separate field-only calibration analysis was also considered because confidence reliability under naturally acquired field imagery is particularly relevant to the intended application. However, the retained prediction-level records do not preserve a sufficiently auditable linkage between each individual test probability vector and its field or PlantVillage source identifier. Field-specific ECE, NLL, and Brier scores were therefore not reconstructed retrospectively from the aggregate results. This avoids introducing source-specific calibration estimates that cannot be verified against the actual test predictions. Accordingly, the present calibration results should be interpreted as aggregate internal mixed-source calibration and not as evidence of calibration under an independently shifted field domain.

3.4 Receiver Operating Characteristic Analysis

The discriminative capability of MultiBranchFusionNet was further assessed using one-versus-rest receiver operating characteristic curves. The per-class ROC-AUC values, rounded to two decimal places, were as follows: Grey Leaf Spot, 1.00; Bacterial Spot, 1.00; Early Blight, 1.00; Late Blight, 1.00; Leaf Mold, 1.00; Septoria Leaf Spot, 1.00; Spider Mites, 1.00; Target Spot, 1.00; Yellow Leaf Curl Virus, 1.00; and Healthy, 1.00. The corresponding ROC curves and per-class AUC values are presented in Figure 6.

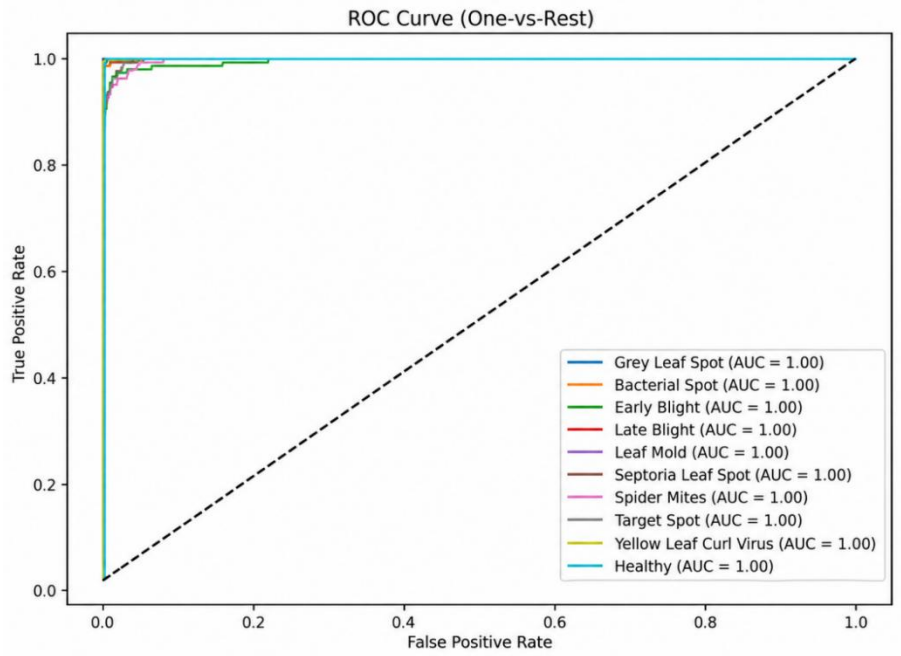


Figure 6. One-versus-rest ROC curves for the ten disease classes. Per-class AUC values, rounded to two decimal places, are reported in the legend.

All evaluated disease categories achieved area under the curve values approaching 1.00, with the individual ROC curves concentrated near the upper-left region of the graph. This distribution reflects high true-positive rates and correspondingly low false-positive rates sustained across multiple classification thresholds, indicating strong class separability throughout the probability space.

Despite the near-perfect AUC values recorded, these findings should be interpreted alongside the confusion matrix results discussed in Section 3.2. While the ROC analysis indicates strong ranking capability across all classes, the confusion matrix evidence indicates that limited misclassification does occur among visually similar disease groups. This distinction is important, as AUC metrics aggregate performance across all decision thresholds and may obscure localised classification difficulties that become apparent only under fixed-threshold evaluation conditions.

Although the near-unity ROC-AUC values support strong discriminative capability within the evaluated dataset, they should not be interpreted as evidence of universally perfect classification performance under other deployment conditions. Model performance may vary when applied to images acquired under substantially different environmental conditions, imaging devices, or disease presentations that were not fully represented during training. Consequently, the reported ROC-AUC values should be viewed as evidence of strong performance within the scope of the experimental evaluation conducted in this study.

3.5 Training Dynamics and Convergence Behaviour

The final MultiBranchFusionNet training run completed the configured maximum of 100 epochs. Although training continued to the configured 100-epoch limit, the checkpoint used for final test evaluation was obtained at epoch 20, where validation categorical accuracy reached its highest recorded value of 0.96888. The corresponding training and validation histories for loss, categorical accuracy, precision, recall, F1-score, and AUC are consolidated in Figure 7. The EarlyStopping callback monitored val_accuracy, while the compiled model logged validation accuracy as val_categorical_accuracy; consequently, the callback did not activate and training continued to the configured 100 epochs. ModelCheckpoint independently monitored val_categorical_accuracy and retained the epoch-20 checkpoint used for final test evaluation.

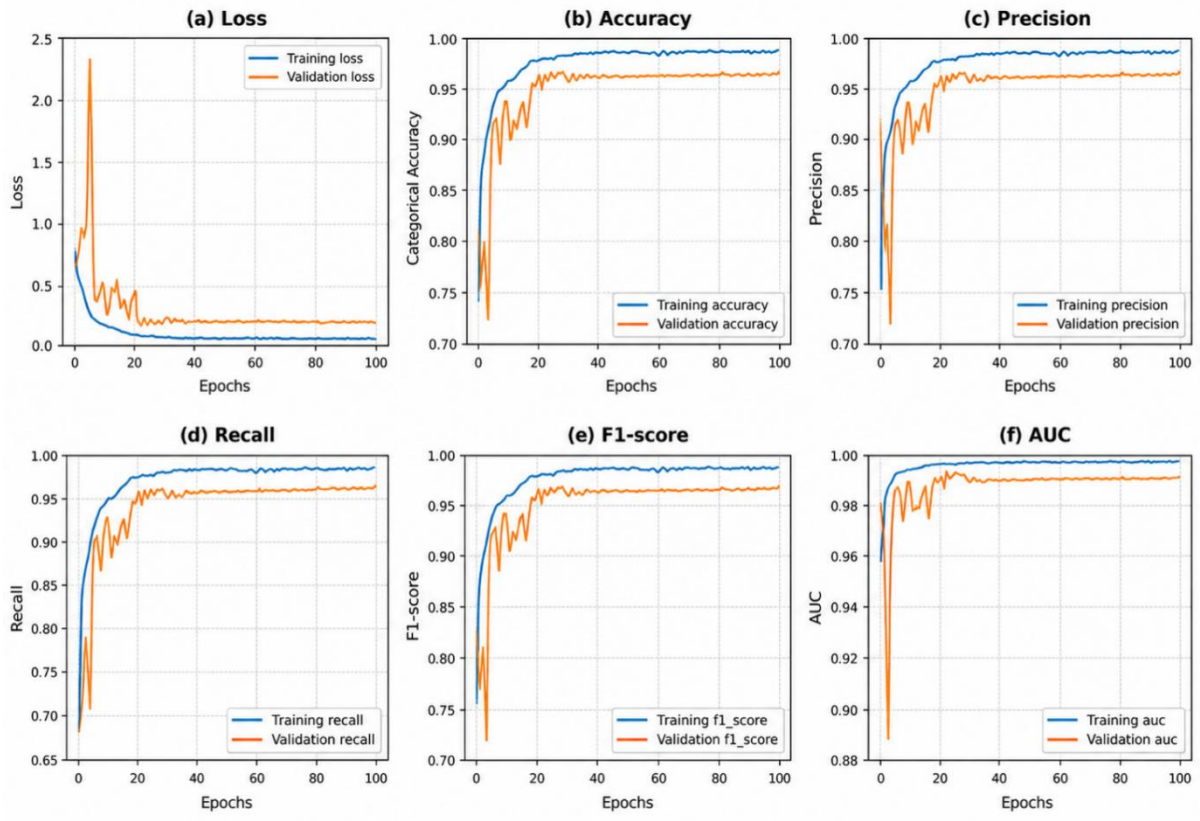


Figure 7. Multi-panel training histories for MultiBranchFusionNet across the complete 100-epoch training run: (a) loss, (b) accuracy, (c) precision, (d) recall, (e) F1-score, and (f) AUC. Final test evaluation used the epoch-20 checkpoint selected by the highest recorded validation categorical accuracy, not the epoch-100 model.

As shown in Figure 7, the model exhibited rapid improvement during the early training epochs, followed by comparatively stable validation behaviour. Training and validation accuracy, precision, recall, F1-score, and AUC converged to consistently high values, while the loss curves showed a progressive decline. Validation loss remained higher than training loss after convergence, indicating some separation between optimisation on the training data and performance on the validation subset. Nevertheless, the validation classification metrics remained comparatively stable throughout the later epochs. These learning histories characterise convergence within the evaluated development data and should not be interpreted as evidence of external generalisation.

3.6 Exploratory Single-Run Ablation Study

An exploratory ablation analysis was used to examine the recorded performance of four architectural configurations during model development: the MobileNetV2 backbone alone, MobileNetV2 with SE attention, a dual-branch configuration combining MobileNetV2 and the custom convolutional pathway without SE attention, and the complete MultiBranchFusionNet architecture containing both the SE mechanism and dual-branch feature fusion. The available results correspond to one recorded run for each configuration and are therefore interpreted descriptively rather than as uncertainty-qualified estimates of the independent effect of each architectural component.

The recorded single-run results are summarised in Table 9. MobileNetV2 alone achieved an accuracy of 92.70%, while the MobileNetV2 + SE configuration recorded 94.10%. The dual-branch configuration without SE attention recorded 95.40%, and the complete MultiBranchFusionNet configuration achieved 97.28%. These values correspond to descriptive differences of 1.40 percentage points between MobileNetV2 and MobileNetV2 + SE, 2.70 percentage points between MobileNetV2 and the dual-branch configuration without SE, and 4.58 percentage points between MobileNetV2 and the complete architecture. Precision, recall, and F1-score showed a similar numerical progression across the recorded configurations.

Table 9. Exploratory single-run ablation results for the MultiBranchFusionNet architecture.

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MobileNetV2 Backbone Only	92.70	93.50	92.00	93.00
MobileNetV2 + SE Attention	94.10	94.30	93.80	94.00
Dual-Branch Fusion without SE Attention	95.40	95.20	95.10	95.10
Proposed MultiBranchFusionNet	97.28	96.75	96.69	96.66

The observed numerical progression is compatible with, but does not independently establish, the architectural rationale that channel recalibration and the task-specific convolutional pathway may provide complementary information. However, the configurations were not evaluated across repeated random seeds, and run-to-run uncertainty is therefore unavailable. In addition, complete configuration-specific measurements of parameter count, GFLOPs, model size, and inference latency were not retained for all four variants. The observed differences cannot consequently be attributed exclusively to the SE mechanism or feature fusion, because changes in model capacity may also contribute. The ablation results are therefore treated as exploratory design-stage evidence rather than as a statistically controlled estimate of component-specific effects.

Values in Table 9 represent one recorded run per architectural configuration and are presented descriptively. Repeated-seed mean ± standard deviation, paired uncertainty estimates, and complete configuration-specific parameter, GFLOP, model-size, and latency measurements were not available for all variants. The differences should therefore not be interpreted as statistically isolated component effects.

3.7 Five-Fold Cross-Validation Analysis

A five-fold cross-validation analysis was conducted on the development portion of the dataset comprising the original training and validation subsets (n = 18,220), while the fixed test subset of 2,025 images remained excluded from the cross-validation analysis. MultiBranchFusionNet and MobileNetV2 were compared using the corresponding recorded fold-wise accuracy results. Because the training portions of conventional cross-validation folds overlap, the five fold-wise observations are not statistically independent. In addition, complete per-image fold assignments, fold-wise class inventories, and acquisition-group mappings were not retained for retrospective verification. The cross-validation results are therefore interpreted as a supplementary descriptive assessment of model behaviour within the development data rather than as independent group-level validation.

Across the five recorded folds, MultiBranchFusionNet achieved accuracies of 96.80%, 96.15%, 97.50%, 96.40%, and 97.10%, corresponding to a mean of 96.79% and a standard deviation of 0.54 percentage points. MobileNetV2 recorded fold accuracies of 93.45%, 92.80%, 94.10%, 93.20%, and 93.85%, with a mean of 93.48% and a standard deviation of 0.52 percentage points. MultiBranchFusionNet recorded the higher accuracy in each of the five matched folds. The resulting paired differences were 3.35, 3.35, 3.40, 3.20, and 3.25 percentage points, yielding a mean recorded difference of 3.31 percentage points. These values describe the observed development-set fold pattern and are not interpreted as five independent replications of the experiment. The fold-wise results and corresponding paired differences are presented in Table 10.

Table 10. Recorded matched five-fold cross-validation results obtained from the development set.

Outer fold	MobileNetV2 accuracy (%)	MultiBranchFusionNet accuracy (%)	Paired difference (percentage points)
Fold 1	93.45	96.80	3.35
Fold 2	92.80	96.15	3.35
Fold 3	94.10	97.50	3.40
Fold 4	93.20	96.40	3.20
Fold 5	93.85	97.10	3.25
Mean ± standard deviation	93.48 ± 0.52	96.79 ± 0.54	3.31 ± 0.08
95% confidence interval	92.84-94.12	96.12-97.46	3.21-3.41

Note: Values represent five matched development-set folds. The 95% CIs were calculated from the five recorded fold-wise values using the Student's t distribution. Because the cross-validation training sets overlap, the intervals should be interpreted with caution and should not be regarded as confidence intervals derived from five independent experimental replications.

The paired accuracy differences ranged from 3.20 to 3.40 percentage points, with a mean of 3.31 percentage points and a sample standard deviation of 0.08 percentage points. The corresponding 95% CI for the mean paired difference was 3.21-3.41 percentage points. Although this interval indicates limited variation among the five recorded fold-wise differences, it should be interpreted cautiously because the cross-validation training sets overlap and the fold-wise observations are therefore not statistically independent.

The numerical calculations underlying the supplementary paired comparison were verified from the five recorded fold differences. A two-sided paired Student's t-test yielded $t(4) = 90.09$ with $p = 9.10 \times 10^{-8}$, whereas the exact two-sided Wilcoxon signed-rank test yielded $W = 0$ with $p = 0.0625$. The discordant p-values reinforce the limited inferential value of formal significance testing based on only five non-independent fold-wise comparisons; neither test is therefore used to support a formal claim of model superiority. None of the five paired differences was zero. The two equal differences of 3.35 percentage points were assigned tied midranks in the signed-rank calculation, and the exact two-sided probability was obtained from the possible sign assignments of the five non-zero paired differences. With five non-zero paired differences all occurring in the same direction, 0.0625 is the minimum attainable exact two-sided probability.

Neither statistical test is treated as definitive evidence of comparative superiority. The paired *t*-test is based on only five fold-wise observations whose training sets overlap, while the Wilcoxon signed-rank test has limited significance resolution at this sample size. Greater emphasis is therefore placed on the unstandardised mean difference, its 95% CI, and the consistency of the observed fold-wise differences rather than on either null-hypothesis significance test considered in isolation. The paired descriptive statistics and supplementary statistical analyses are summarised in Table 11.

Table 11. Paired descriptive statistics and supplementary statistical analyses of the recorded fold-wise accuracies.

Statistical measure	Result
Paired differences (percentage points)	3.35, 3.35, 3.40, 3.20, 3.25
Mean paired difference	3.31
Standard deviation of paired differences	0.08
95% CI for the mean paired difference	3.21-3.41
Paired t-statistic	90.09
Degrees of freedom	4
Paired t-test p-value	9.10×10^{-8}
Exact two-sided Wilcoxon signed-rank test	$W = 0$, $p = 0.0625$

Note: The paired statistical tests are reported as supplementary sensitivity analyses. The five fold-wise observations are not independent because the corresponding cross-validation training sets overlap. For the Wilcoxon signed-rank calculation, there were no zero differences and tied absolute differences were assigned midranks. Neither test is interpreted as definitive evidence of model superiority.

The five recorded fold-wise comparisons therefore show a consistent directional difference within the evaluated development data, with MultiBranchFusionNet recording higher accuracy than MobileNetV2 in each fold. Nevertheless, the analysis does not constitute independent repeated validation because the cross-validation training sets overlap and complete fold-level acquisition-group assignments were not retained for retrospective verification. Stronger comparative validation will require repeated group-aware cross-validation or repeated independent splits defined prospectively at the plant, farm, acquisition-session, or source level, with complete fold membership and class distributions retained for reproducibility.

3.8 Contextual Comparison with Recent Lightweight and Attention-Based Methods

To provide a broader contextual comparison with recent lightweight and attention-based plant disease classification methods, Table 12 summarises reported performance and computational characteristics from recent studies, including work published between 2024 and 2026. The comparison includes accuracy, precision, recall, F1-score, parameter count, FLOPs, and dataset and evaluation characteristics where reported. Values were reproduced from the respective publications without recalculation. Where a study did not report a particular metric, the entry is marked as NR rather than inferred from other results. Inference time is not included as a direct cross-study metric because the reviewed studies used different hardware platforms, input resolutions, batch sizes, runtimes, and implementation environments, making direct latency comparisons unreliable.

Table 12. Cross-study comparison of recent lightweight and attention-based plant disease classification models.

Study	Dataset and evaluation setting	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Parameters (M)	FLOPs (G)
[22]	PlantVillage tomato dataset; 14,531 images; 10 classes; controlled image-acquisition conditions	95.60	95.52	95.50	95.40	86.60	NR
[28]	Public PlantVillage-derived tomato dataset; 10 classes; controlled conditions; test evaluation with tenfold validation	99.11	99.00	99.00	NR	NR	NR
[42]	Self-collected tomato dataset; 6,001 original images expanded through augmentation; 4 classes; simulated complex environmental conditions; 80:20 split	97.84	NR	NR	NR	0.418	NR
[43]	PlantVillage dataset; approximately 55,000 images; 38 classes; 14 crop species; controlled conditions; test evaluation with five-fold validation	98.70	98.30	99.10	98.70	2.40	NR
[44]	Four public plant disease datasets, including PlantVillage and Xinong Apple; separate controlled benchmark evaluations	97.08-99.99	NR	NR	NR	NR*	NR
[45]	PlantVillage grapevine dataset; 4 classes; controlled conditions; five-fold cross-validation mean	98.62	98.68	98.65	98.63	0.173	NR
[48]	PlantVillage dataset; 54,305 images; 38 classes; 14 crop species; controlled conditions; 80:10:10 split	99.79	99.70	99.72	99.71	3.34	NR
[49], ShuffleNetV2	PlantVillage tomato test set; controlled conditions	98.29	NR	NR	97.70	1.26	0.15
[49], MobilePlantViT	PlantVillage tomato test set; controlled conditions; lightweight model evaluation	99.40	NR	NR	99.31	0.82	0.60
This Study	Hybrid field-acquired and PlantVillage tomato dataset; 20,245 images; 10 classes; heterogeneous backgrounds and acquisition conditions; 70:20:10 split	97.28	96.75	96.69	96.66	2.96	1.63

NR: not reported in the corresponding publication. *Deng et al. [44] reported a parameter volume equivalent to 37.5% of InceptionV4 but did not provide an exact parameter count in millions. All reported values are reproduced from the respective publications without recalculation. Precision, recall, and F1-score averaging conventions may differ where the original publication does not specify the averaging method. Because the compared studies used different datasets, crop species, disease classes, image-acquisition conditions, partitioning strategies, hardware platforms, and software environments, Table 12 should be interpreted as a contextual comparison rather than a common-benchmark ranking.

Recent lightweight and attention-based models have continued to improve plant disease classification while reducing computational requirements. CNN-SEEIB [48] achieved 99.79% accuracy, 99.71% F1-score, and 3.34 million parameters on the PlantVillage dataset, providing a recent reference point for lightweight plant disease classification under controlled image-acquisition conditions.

Hoang et al. [49] further illustrate the trade-off between diagnostic performance and computational cost. Their 2026 evaluation reported 99.40% accuracy and 99.31% macro-F1 for MobilePlantViT with 0.82 million parameters and 0.60 GFLOPs, while ShuffleNetV2 achieved 98.29% accuracy and 97.70% macro-F1 with 1.26 million parameters and 0.15 GFLOPs. These results provide useful computational reference points for MultiBranchFusionNet, although the models were evaluated on the PlantVillage dataset under a different experimental setting and therefore should not be interpreted as direct performance benchmarks against the present study. Domain-adaptation methods have also been proposed to improve cross-domain robustness [50].

The comparison illustrates how dataset composition and image-acquisition conditions can influence reported performance across studies. LWDSC-SA [43] and CNN-SEEIB [48], for example, achieved accuracies above 98% on PlantVillage. Although these studies involved 38 classes, PlantVillage images are generally captured under relatively uniform backgrounds and controlled imaging conditions. The number of classes alone therefore does not establish that their

classification tasks were more difficult than the present task. Moreover, multi-crop datasets may contain differences in leaf morphology and colour that help distinguish some classes, whereas the present study requires discrimination among visually similar symptoms occurring on the same crop.

The studies by Vithalani and Dabhi [22] and Assaduzzaman et al. [28] are more closely related to the present work because both evaluated 10-class tomato disease classification. Nevertheless, their evaluations were conducted predominantly with controlled public images. The 95.60% accuracy reported for the fine-tuned ViT-Base model in [22] and the 99.11% reported for XSE-TomatoNet [28] therefore provide useful reference points, but neither result constitutes a direct comparison with MultiBranchFusionNet under the same test conditions.

The dataset used by Sun et al. [42] contains self-collected tomato images and is consequently relevant to classification outside laboratory conditions. However, that study considered only four categories and substantially enlarged the original dataset through augmentation. Similarly, ALNet [45] was evaluated on four grapevine classes rather than tomato diseases. DAMSLNet [44] reported accuracies ranging from 97.08% to 99.99% across four public datasets, further illustrating how performance can vary with the crop species, class definitions, dataset size, and evaluation protocol.

Within the present evaluation, MultiBranchFusionNet achieved 97.28% accuracy with macro-precision, macro-recall, and macro-F1-score values of 96.75%, 96.69%, and 96.66%, respectively, on the assembled 10-class mixed-source dataset. These results provide a contextual reference point within the recent literature; however, the studies in Table 12 differ in dataset composition, acquisition conditions, class definitions, partitioning procedures, optimisation, and hardware. The comparison is therefore descriptive rather than a common-benchmark ranking. A controlled same-dataset evaluation of recent lightweight multi-branch and attention-based architectures remains necessary for direct comparative assessment.

3.9 Comparative Discussion

The overall and macro-averaged classification results indicate that MultiBranchFusionNet maintained strong performance across the evaluated classes rather than deriving its aggregate accuracy solely from the more frequently represented categories. Nevertheless, the lower recall for Spider Mites and the lower F1-scores for Early Blight and Target Spot identify specific categories where visually overlapping symptom patterns continue to present classification difficulty.

The exploratory ablation and development-set fold analyses provide complementary evidence concerning the proposed architecture. The recorded ablation sequence was directionally consistent with the intended roles of SE-based channel recalibration and dual-branch feature fusion, while MultiBranchFusionNet recorded higher accuracy than MobileNetV2 across all five matched development folds. These findings remain descriptive because the ablation configurations were single-run experiments and the cross-validation training folds overlapped; their detailed methodological qualifications are reported in Sections 3.6, 3.7, and 3.10.

Calibration analysis further showed improved probabilistic agreement after validation-fitted temperature scaling across ECE, categorical NLL, and multiclass Brier score. However, these results apply to the internal mixed-source distribution and do not establish calibration stability under independently shifted field conditions. Similarly, combining PlantVillage and field-acquired images increases the acquisition diversity and environmental realism represented within the internal evaluation distribution; however, mixing both sources before image-level partitioning does not constitute source-independent external validation or demonstrate robustness to domain shift. Recent plant-disease work has also examined confidence and uncertainty under real-world field conditions [51].

MultiBranchFusionNet retains a compact architecture-level profile, with 2.96 million parameters, a model size of 15.28 MB, and a nominal computational requirement of 1.63 GFLOPs per forward pass. The FastAPI implementation provides cloud-served inference, while reproducible hardware-latency and direct on-device performance remain to be established.

Taken together, the findings support the effectiveness of the proposed attention-guided asymmetric dual-branch integration within the evaluated mixed-source dataset. The contribution should be understood as an engineering integration and internal empirical evaluation of established pretrained, attention, task-specific convolutional, and feature-fusion components rather than as evidence of universal superiority over recent alternative architectures. The principal limitations and corresponding future directions are consolidated in Section 3.10.

The channel-wise recalibration provided by the SE block does not, by itself, indicate where in an image the evidence supporting a prediction is located. Branch-specific Grad-CAM was therefore used to examine representative correct and incorrect MultiBranchFusionNet predictions from the held-out test subset. The selected visualisations include examples from the field-acquired and PlantVillage image components where source identity was verified, together with failure cases involving Spider Mites, Early Blight, Target Spot, and practically important diseased-to-Healthy predictions. Representative Grad-CAM results are presented in Figure 8.

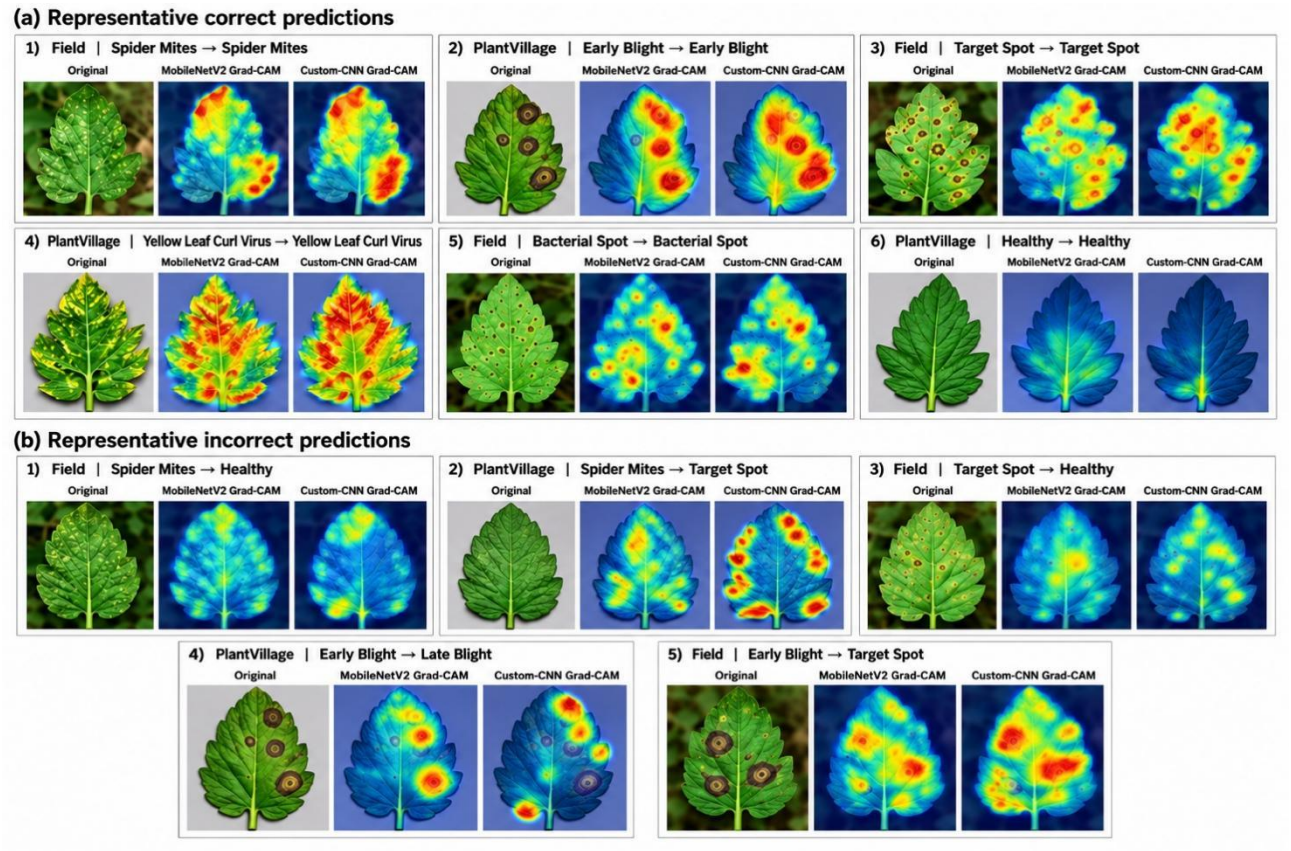


Figure 8. Representative branch-specific Grad-CAM visualisations for correct and incorrect MultiBranchFusionNet predictions. (a) Representative correct predictions, (b) Representative incorrect predictions.

As depicted in Figure 8, original images are shown with Grad-CAM overlays from the terminal MobileNetV2 and custom-CNN convolutional representations. The examples include field-acquired and PlantVillage images and representative cases involving Spider Mites, Early Blight, Target Spot, Yellow Leaf Curl Virus, Bacterial Spot, Healthy leaves, and disease-to-Healthy errors. Warmer regions indicate greater spatial contribution to the predicted class.

Across the representative correct predictions, the Grad-CAM maps were predominantly concentrated on the leaf surface rather than exclusively on the surrounding image background. For disease-positive images, stronger activations generally overlapped visually symptomatic regions, including localised spots, necrotic areas, mottling, chlorotic regions, and broader symptom-bearing portions of the leaf. The activation pattern for Yellow Leaf Curl Virus was spatially broader than for several spot-forming diseases, which is consistent with its more distributed leaf-level appearance. Correct Healthy predictions showed comparatively diffuse activation across the leaf structure rather than concentration on discrete lesion-like regions.

The two branches also exhibited differences in spatial emphasis. The terminal MobileNetV2 maps generally produced broader high-level activation regions, whereas the custom-CNN branch showed more localised responses in several examples. This difference is consistent with their complementary architectural roles: the pretrained pathway supplies transferable high-level representations, while the custom branch learns disease-specific features directly from the tomato leaf dataset. Nevertheless, the visual overlap between Grad-CAM activations and symptomatic regions should not be interpreted as validated biological lesion localisation or as proof that either branch exclusively represents lesions or any single morphological characteristic.

The incorrect predictions provided additional insight into the residual error patterns identified by the confusion matrix. In representative Spider Mites and Target Spot images predicted as Healthy, the activation maps were comparatively diffuse or showed weaker concentration on the visible symptomatic regions than was observed in corresponding correctly classified disease examples. This pattern is of practical concern because the confusion matrix contained ten Spider Mites-to-Healthy errors and five Target Spot-to-Healthy errors. These disease-to-Healthy failures suggest that, for some images, disease-associated visual evidence may not have contributed sufficiently strongly to overcome features associated with the Healthy class.

By contrast, representative disease-to-disease errors retained activation over symptom-bearing leaf regions even when the final disease label was incorrect. For example, misclassified Early Blight cases showed activation associated with visible lesion areas despite assignment to another visually similar disease category. Such cases suggest that some residual errors reflect difficulty in discriminating among overlapping symptom morphologies rather than a complete failure to localise

abnormal leaf regions. The Spider Mites-to-Target Spot example similarly illustrates how localised spot-like evidence can support an incorrect but visually related disease prediction.

Overall, the representative Grad-CAM visualisations provide supportive qualitative evidence that the examined predictions were associated primarily with leaf and symptom-bearing regions, while selected failure cases revealed diffuse attribution and ambiguity among visually overlapping disease patterns. Background contribution cannot be excluded, and the limited qualitative sample does not establish source invariance or causal dependence on the highlighted regions. These limitations and the need for broader source-stratified attribution analysis are addressed in Section 3.10.

3.10 Study Limitations and Future Directions

The findings should first be interpreted within the geographical, acquisition, and partitioning scope of the assembled dataset. The field component comprised 9,312 images collected from 26 farms across Ogun, Oyo, and Ondo States during two acquisition phases, but this coverage does not represent the full range of tomato cultivars, disease strains, climatic conditions, agricultural practices, or seasons encountered elsewhere. All field images were acquired using an iPhone 14 Pro Max, further limiting device diversity. In addition, train, validation, and test partitions were constructed at image level rather than prospectively separated by plant, farm, acquisition session, state, collection phase, or image source. Multiple views could originate from the same plant or acquisition context; therefore, residual dependence between partitions cannot be excluded. Such images may represent correlated observations rather than independent biological samples. If related images occurred across development and evaluation subsets, image-level performance could be optimistic relative to evaluation on entirely unseen plants, farms, acquisition sessions, or sources. The extent of this effect cannot be determined retrospectively, as the required grouping identifiers were not retained. The reported performance should therefore be interpreted as internal mixed-source image-level evaluation rather than independent field-domain generalisation. Accordingly, the present experiment does not establish PlantVillage-to-field transfer, field-to-PlantVillage transfer, or performance under a prospectively held-out source, farm, state, or geographical domain. Future evaluation will employ prospectively recorded acquisition identifiers and group-separated, source-separated, and geographically broader test sets acquired across multiple devices and environmental conditions.

A second limitation concerns reference-label quality and class distribution. Field labels were derived from visual symptom assessment with expert adjudication of uncertain cases rather than laboratory confirmation, and a complete annotation audit trail was not retained. Visually similar symptoms may also arise from nutrient deficiencies, pest damage, chemical injury, environmental stress, or mixed infections; without laboratory confirmation, these alternative causes cannot be excluded completely. The dataset was also imbalanced across classes, while the final training configuration did not use class weighting, balanced mini-batches, or explicit minority-class oversampling. Macro-averaged and class-specific metrics were reported to make this behaviour visible. Future work will incorporate prospectively documented annotation procedures, laboratory confirmation where feasible, and controlled evaluation of alternative imbalance-handling strategies.

The architecture-development analyses also have design limitations. The ablation study contained one recorded run per configuration and did not include repeated-seed or capacity-normalised comparisons. The five-fold development analysis used overlapping training folds and therefore does not constitute five independent experimental replications. Similarly, the preliminary backbone screening was intended to support engineering selection rather than provide an exhaustively optimised head-to-head benchmark. The implemented preprocessing also used uniform [0,1] rescaling without MobileNetV2-specific preprocess_input, and the effect of this choice was not isolated experimentally. This input-scale mismatch may have altered the feature responses of the frozen pretrained backbone relative to its standard ImageNet preprocessing regime; consequently, the present results do not establish whether MobileNetV2-specific preprocessing would improve, reduce, or otherwise alter classification performance. Future comparative work should use matched partitions, repeated random seeds, complete fold records, architecture-specific optimisation, and controlled preprocessing comparisons.

Calibration and explainability findings remain bounded by the internal evaluated distribution. Validation-fitted temperature scaling improved ECE, categorical NLL, and multiclass Brier score, but source-specific and independently shifted field-domain calibration were not established. Grad-CAM provided qualitative spatial evidence for representative predictions but does not establish causal dependence on highlighted regions or demonstrate invariance to background and source-related cues. Future studies should therefore assess calibration on independently collected field data and extend explainability evaluation using larger source-stratified attribution sets, complementary explanation methods, and quantitative localisation where reliable lesion annotations are available.

Finally, the parameter count, model size, and nominal FLOP requirement indicate a compact architecture but do not establish operational performance on resource-constrained hardware. The implemented FastAPI service performs cloud-served rather than local on-device inference, and a reproducible device-level latency benchmark was not available. Peak inference memory, energy consumption, thermal behaviour, model-conversion performance, and quantisation effects were likewise not evaluated. Future deployment studies will therefore benchmark the framework directly on representative mobile and embedded platforms while measuring latency distributions, memory use, energy consumption, conversion and quantisation behaviour, and reliability under routine field conditions.

4. Conclusion

This study developed MultiBranchFusionNet, a lightweight attention-guided dual-branch framework for classifying ten tomato leaf conditions using 20,245 field-acquired and publicly sourced images. For the field-acquired component, the class labels represent symptom-based expert assessments rather than laboratory-confirmed pathogen diagnoses. The architecture combines a frozen ImageNet-pretrained MobileNetV2 pathway enhanced with SE attention and a custom convolutional pathway through feature-level fusion. Under the evaluated mixed-source image-level protocol, the model achieved 97.28% accuracy, with macro-precision, macro-recall, and macro-F1-score values of 96.75%, 96.69%, and 96.66%, respectively. The image-level partitioning strategy and absence of complete plant- and acquisition-group identifiers mean that these values quantify internal image-level classification rather than plant-, farm-, acquisition-session-, or source-independent generalisation. With 2.96 million parameters, a model size of 15.28 MB, and a nominal computational requirement of 1.63 GFLOPs, the framework also retains a compact architecture-level profile. Exploratory ablation, development-fold, calibration, and Grad-CAM analyses provided complementary evidence concerning model behaviour, subject to the methodological constraints consolidated in Section 3.10. Overall, the findings support the effectiveness of the proposed attention-guided asymmetric dual-branch integration within the evaluated mixed-source dataset, but they should not be interpreted as evidence of independent plant-, farm-, source-, geographical-, device-, or deployment-level generalisation. Future work will therefore prioritise prospectively grouped external field validation and direct benchmarking on representative mobile and embedded hardware.

Acknowledgements

Not Applicable.

Funding

This research received no external funding.

Ethics Statement

This study involved the acquisition and analysis of non-human plant images and did not involve the recruitment of human participants, collection of human biological material, or processing of human-subject data. Formal human-subject ethical approval was therefore not sought for the plant-image collection. Before image acquisition at each of the 26 participating farms, access to the farm and permission to photograph tomato plants for research purposes were obtained from the farm manager or other responsible personnel present at the location.

The image-acquisition protocol was directed at tomato plants and leaves rather than people, and no names, contact information, health information, or other personal identifiers of farm personnel were collected as part of the dataset. Exact farm coordinates and identifying farm details are not reported publicly. Any field images subsequently provided to external researchers will be reviewed for incidental identifying content and location-identifying metadata, where present, before transfer so that the privacy of participating farms and personnel is protected.

The publicly sourced component was obtained from the PlantVillage Tomato Leaf Dataset hosted on Kaggle and cited in this manuscript [46]. The images were accessed and used under the conditions presented by the source repository. Because the specific Kaggle listing does not identify a named reuse licence, this study does not assert broader redistribution rights beyond those provided by the repository. The field-acquired component remains under the stewardship of the authors and may be made available for legitimate research purposes upon reasonable request to the corresponding author, subject to appropriate protection of farm privacy and identifying location information.

Data Availability Statement

The publicly sourced tomato leaf images used in this study are accessible through the PlantVillage Tomato Leaf Dataset hosted on Kaggle at https://www.kaggle.com/datasets/charuchaudhry/plantvillage-tomato-leaf-dataset, subject to the access and reuse conditions presented by the repository. The field-acquired images collected from farms in Ogun, Oyo, and Ondo States, Nigeria, are available from the corresponding author upon reasonable request for legitimate research purposes, subject to protection of farm privacy and removal of identifying location metadata or incidental identifying content where applicable. The implementation code has not yet been deposited in a public repository; therefore, no public code-repository link is claimed in the present study.

Author Contributions

Olayiwola Charles Adesoba: Conceptualization, methodology, investigation, visualisation, writing and editing. Waliu Olalekan Apena: Conceptualization, supervision, and validation. Folasade Mojisola Dahunsi: Conceptualization, supervision, methodology, and validation. Ayodele Martins Ajayi: Investigation and data analysis, and supervision.

Conflict of Interest

The authors declare that they have no conflict of interest.

Generative AI Statement

Generative AI tools assisted with language editing and grammatical improvement. The authors retain responsibility for the study design, experimental data, reported results, and interpretations.

References

[1] Akotowanou OCA, Adjou ES, Olubi AB, Kougblenou SD, Ahoussi ED, Sohounhloué DC. The tomato (*Solanum lycopersicum L.*) in community development: An overview focuscent achievements and future prospective. International Journal of Frontiers in Biology and Pharmacy Research, 2022, 3(2), 8-16. DOI: 10.53294/ijfbpr.2022.3.2.0061

[2] Collins EJ, Bowyer C, Tsouza A, Chopra M. Tomatoes: An extensive review of the associated health impacts of tomatoes and factors that can affect their cultivation. Biology, 2022, 11(2), 239. DOI: 10.3390/biology11020239

[3] Khan UM, Sevindik M, Zarrabi A, Nami M, Ozdemir B, Kaplan DN, et al. Lycopene: Food sources, biological activities, and human health benefits. Oxidative medicine and cellular longevity, 2021, 2021(1), 2713511. DOI: 10.1155/2021/2713511

[4] Mohammed KO, Alim AS, Abubakar ZU, Ibrahim T. Impact analysis of tomato contract farming on farmers welfare in Kano State, Nigeria. Journal of Entomology and Agronomy Studies, 2026, 11(1), 23-40. DOI: 10.70382/bjeas.v11i1.036

[5] Modiba MP, Bell T, Glick B, Babalola OO. *Ralstonia solanacearum* and *Xanthomonas perforans* as causal agents of bacterial disease of tomato. MicrobiologyOpen, 2025, 14(6), e70195. DOI: 10.1002/mbo3.70195

[6] Oladipo EK, Oladipo B, Ojumu TV, Caleb OJ. Postharvest losses of tomato: Causes, spoilage mechanisms, and advances in conventional and emerging technologies for preservation. Food Research International, 2026, 227, 118247. DOI: 10.1016/j.foodres.2025.118247

[7] Khan M, Gulan F, Arshad M, Zaman A, Riaz A. Early and late blight disease identification in tomato plants using a neural network-based model to augmenting agricultural productivity. Science Progress, 2024, 107(3), 368504241275371. DOI: 10.1177/00368504241275371

[8] Panthee DR, Pandey A, Paudel R. Multiple foliar fungal disease management in tomatoes: A comprehensive approach. International Journal of Plant Biology, 2024, 15(1), 69-93. DOI: 10.3390/ijpb15010007

[9] Jonathan SG, Okunola TT, Nwaokolo VM, Omeonu FC, Adegoke BA, Salaam SM. Challenges confronting plant disease management in Nigeria. Journal of Agriculture and Aquaculture, 2025, 7(1), 1-7. DOI: 10.5281/zenodo.14955777

[10] Lahlali R, Taoussi M, Laasli S, Gachara G, Ezzouggari R, Belabess Z, et al. Effects of climate change on plant pathogens and host-pathogen interactions. Crop and Environment, 2024, 3(3), 159-170. DOI: 10.1016/j.crope.2024.05.003

[11] Zhou J, Zhang X, Qu Z, Zhang C, Wang F, Gao T, et al. Progress in research on prevention and control of crop fungal diseases in the context of climate change. Agriculture, 2024, 14(7), 1108. DOI: 10.3390/agriculture14071108

[12] Sawadogo S, Tiendrebeogo F, Tibiri EB, Name PE, Djigma F, Traoré L, et al. Artificial intelligence tool for cassava viral diseases diagnosis using participatory surveillance in Burkina Faso. Frontiers in Sustainable Food Systems, 2025, 9, 1597039. DOI: 10.3389/fsufs.2025.1597039

[13] Adesoba OC. A deep learning approach for disease detection in maize crops. IT Journal Research and Development, 2025, 10(1), 53-66. DOI: 10.25299/itjrd.2025.22875

[14] Das A, Pathan F, Jim JR, Kabir MM, Mridha M. Deep learning-based classification, detection, and segmentation of tomato leaf diseases: A state-of-the-art review. Artificial Intelligence in Agriculture, 2025, 15(2), 192-220. DOI: 10.1016/j.aiia.2025.02.006

[15] Ghimire S, Lamsal R, Sankuratri A, Lakkarsu P, Vutla S. Artificial intelligence-driven plant disease detection and diagnosis: A comprehensive review of deep learning approaches, multimodal sensing technologies, and future perspectives in precision agriculture. The Plant Pathology Journal, 2026, 42(2), 121-137. DOI: 10.5423/PPJ.RW.01.2026.0004

[16] Mamo TY. Plant pathology: Crossing the boundaries with novel approaches and perspectives for scientists. Pathology and Laboratory Medicine, 2024, 8(1), 1-8. DOI: 10.11648/j.plm.20240801.11

[17] Adesoba OC, Apena WO, Dahunsi FM, Ajayi AM. Hybrid deep learning models for efficient detection and classification of tomato plant leaf diseases. Journal of Artificial Intelligence and System Modelling, 2026, 4(1), 18-42. DOI: 10.22034/jaism.2026.552056.1164

[18] Gunasekaran H, Rajkumar S, Kirubhadharsini BL. Lightweight deep learning for tomato disease detection: Trends, challenges, and edge AI perspectives. Frontiers in Plant Science, 2026, 16, 1737208. DOI: 10.3389/fpls.2025.1737208

[19] Tian K, Zeng J, Song T, Li Z, Evans A, Li J. Tomato leaf diseases recognition based on deep convolutional neural networks. Journal of Agricultural Engineering, 2023, 54(1), 1432. DOI: 10.4081/jae.2022.1432

[20] Dhanya R, Mythili S, Rexeena X, Devadas D. Detection of tomato leaf diseases using deep learning and spatial attention mechanisms. International Journal of Research and Scientific Innovation, 2025, 12(7), 1144-1153. DOI: 10.51244/IJRSI.2025.120700118

[21] Jelali M. Deep learning networks-based tomato disease and pest detection: A first review of research studies using real field datasets. Frontiers in Plant Science, 2024, 15, 1493322. DOI: 10.3389/fpls.2024.1493322

[22] Vithalani SK, Dabhi VK. A novel deep learning approach for tomato leaf disease detection using optimized CNN architecture. Journal of Computer Science, 2026, 22(1), 47-60. DOI: 10.3844/jcssp.2026.47.60

[23] Mallick MT, Banerjee S, Thakur N, Saha HN, Chakrabarti A. Evaluation of state-of-the-art deep learning techniques for plant disease and pest detection. Computers, Materials & Continua, 2025, 85(1), 121-180. DOI: 10.32604/cmc.2025.065250

[24] Mohanty SP, Hughes DP, Salathé M. Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 2016, 7, 1419. DOI: 10.3389/fpls.2016.01419

[25] Ferentinos KP. Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 2018, 145, 311-318. DOI: 10.1016/j.compag.2018.01.009
[26] Too EC, Yujian L, Njuki S, Yingchun L. A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 2019, 161, 272-279. DOI: 10.1016/j.compag.2018.03.032
[27] Deb N, Rahman T. An efficient VGG16-based deep learning model for automated potato pest detection. Smart Agricultural Technology, 2025, 12, 101409. DOI: 10.1016/j.atech.2025.101409
[28] Assaduzzaman M, Bishshash P, Nirob MAS, Marouf AA, Rokne JG, Alhajj R. XSE-TomatoNet: An explainable AI-based tomato leaf disease classification method using EfficientNetB0 with squeeze-and-excitation blocks and multi-scale feature fusion. MethodsX, 2025, 14, 103159. DOI: 10.1016/j.mex.2025.103159
[29] Zhao J, Xu L, Ma Z, Li J, Wang X, Liu Y, et al. A review of plant leaf disease identification by deep learning algorithms. Frontiers in Plant Science, 2025, 16, 1637241. DOI: 10.3389/fpls.2025.1637241
[30] Desai J, Ganatra A. Comprehensive study of various machine learning models for plant disease detection: Analysis of deep models on tomato plant. International Journal of Innovative Research and Scientific Studies, 2025, 8(3), 4962-4976. DOI: 10.53894/ijirss.v8i3.7697
[31] Barbedo JGA. Plant disease identification from individual lesions and spots using deep learning. Biosystems Engineering, 2019, 180, 96-107. DOI: 10.1016/j.biosystemseng.2019.02.002
[32] Song X, Yan L, Liu S, Gao T, Han L, Jiang X, et al. Agricultural image processing: Challenges, advances, and future trends. Applied Sciences, 2025, 15(16), 9206. DOI: 10.3390/app15169206
[33] Vijayakanthan G, Nasser M, Algbear AF, Alosaimi HM, Ibrahim AO, Ravi V. Explainable and lightweight deep learning models for tea leaf disease detection: A systematic review and future research directions. Smart Agricultural Technology, 2026, 14, 102042. DOI: 10.1016/j.atech.2026.102042
[34] Poonam, Nayak SK, Mishra R. A deep learning approach for optimizing edge computing for real-time IoT applications. International Journal of Research and Review in Applied Science, Humanities, and Technology, 2025, 2(2), 43-51. DOI: 10.71143/w40tra50
[35] Li L, Liu CZ, Cao K, Zhou Z, Alrowaily M. Artificial intelligence of things for smart agriculture with optimal embedded edge intelligence. Computers & Electrical Engineering, 2026, 133, 111039. DOI: 10.1016/j.compeleceng.2026.111039
[36] Tiwari V, Joshi RC, Dutta MK. Dense convolutional neural networks based multiclass plant disease detection and classification using leaf images. Ecological Informatics, 2021, 63, 101289. DOI: 10.1016/j.ecoinf.2021.101289
[37] Abetu MT, Abebe TU, Tune KK. Efficient and robust deep learning for field-level wheat disease classification on resource-constrained devices. Smart Agricultural Technology, 2026, 14, 102091. DOI: 10.1016/j.atech.2026.102091
[38] Karimanzira D. Context-aware tomato leaf disease detection using deep learning in an operational framework. Electronics, 2025, 14(4), 661. DOI: 10.3390/electronics14040661
[39] Ashurov AY, Al-Gaashani MSAM, Samee NA, Alkanhel R, Atteia G, Abdallah HA, et al. Enhancing plant disease detection through deep learning: A depthwise CNN with squeeze and excitation integration and residual skip connections. Frontiers in Plant Science, 2025, 15, 1505857. DOI: 10.3389/fpls.2024.1505857.
[40] Lu X, Yang M, Chen Y, Xiong S, Lu X. Multibranch fusion-based feature enhance for remote-sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63, 1-17. DOI: 10.1109/TGRS.2025.3546638
[41] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8), 2011-2023. DOI: 10.1109/TPAMI.2019.2913372
[42] Sun W, Xu Z, Xu K, Ru L, Yang R, Wang R, et al. Ultra-lightweight tomatoes disease recognition method based on efficient attention mechanism in complex environment. Frontiers in Plant Science, 2025, 15, 1491593. DOI: 10.3389/fpls.2024.1491593
[43] Batool A, Kim J, Byun YC. A compact deep learning approach integrating depthwise convolutions and spatial attention for plant disease classification. Plant Methods, 2025, 21, 48. DOI: 10.1186/s13007-025-01325-4
[44] Deng L, Qin J, Li K, Zhu J, Wang Z. DAMSLNet: Dual-attention multi-scale lightweight network for plant disease classification. Engineering Applications of Artificial Intelligence, 2025, 159, 111590. DOI: 10.1016/j.engappai.2025.111590
[45] Adithiyaa D, Lakshhmi Narayanan T, Prusty MR. An attention-augmented lightweight convolutional framework for fine-grained plant leaf disease classification. Frontiers in Plant Science, 2026, 17, 1762956. DOI: 10.3389/fpls.2026.1762956
[46] Chaudhry C. PlantVillage tomato leaf dataset. Kaggle. 2021. Available from: https://www.kaggle.com/datasets/charuchaudhry/plantvillage-tomato-leaf-dataset (accessed on 22 June 2025).
[47] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, 2017, 70, 1321-1330. Available from https://proceedings.mlr.press/v70/guo17a.html (accessed on 12 July 2025).
[48] Karthikeyan S, Charan R, Narayanan S, Jani Anbarasi L. Enhanced plant disease classification with attention-based convolutional neural network using squeeze and excitation mechanism. Frontiers in Artificial Intelligence, 2025, 8, 1640549. DOI: 10.3389/frai.2025.1640549
[49] Hoang T, Bui V, Nguyen V, Doan D, Dang H, Pham A. A comprehensive evaluation of lightweight deep learning models for tomato disease classification on edge computing environments. Scientific Reports, 2026, 16, 12320. DOI: 10.1038/s41598-026-42439-6
[50] Tunio MH, Li JP, Zeng X, Awais A, Shah SA, Shaikh H-U, et al. Advancing plant disease classification: A robust and generalized approach with transformer-fused convolution and Wasserstein domain adaptation. Computers and Electronics in Agriculture, 2024, 227, 109574. DOI: 10.1016/j.compag.2024.109574
[51] Ergün E, Okumus H, Cinar OE, Batan N. Confidence and uncertainty aware deep learning for reliable grape leaf disease diagnosis under real world field conditions. Frontiers in Plant Science, 2026, 17, 1839652. DOI: 10.3389/fpls.2026.1839652