Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 2, 2025, 227-240.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Multi-Head Self-Attention Fusion Network for Enhanced Multi-Class Crop Disease Classification

Thomas Kinyanjui Njoroge[1*], Kelvin Mugoye[2] and Rachael Kibuku[2]

1. Computer Science & Informatics Department, School of Pure and Applied Science, Karatina University, Kenya.
2. Software Development & Information Systems, School of Technology, KCA University, Nairobi, Kenya.

Article Info

Article History:
Received 31 January 2025
Revised 06 February 2025
Accepted 04 March 2025

DOI:10.22044/jadm.2025.15689.2687

Keywords:
Crop Disease Detection, Vision Transformers, Convolutional Neural Networks, Multimodal Fusion, Deep Learning.

**Corresponding author: tnjoroge@karu.ac.ke (T. Kinyanjui).*

Abstract

This paper presents a Multi-Head Self-Attention Fusion Network (MHSA-FN) for real-time crop disease classification, addressing key limitations in existing models, including suboptimal feature extraction, inefficient feature recalibration, and weak multi-scale fusion. Unlike prior works that rely solely on CNNs or transformers, MHSA-FN integrates MobileNetV2, EfficientNetV2, and Vision Transformers (ViTs) with a structured multi-level attention framework for enhanced feature learning. A gated fusion mechanism and a Multi-scale Fusion Module (MSFM) optimize local texture details and global spatial relationships. The model was trained on a combined dataset of PlantVillage and locally collected images, improving adaptability to real-world conditions. It achieved 98.66% training accuracy and 99.0% test accuracy across 76 disease classes, with 99.34% precision, 99.01% recall, and 99.04% F1 score. McNemar's test ($p = 0.125$) and Bayesian superiority probability (0.851) validated its robustness. Confidence variance analysis (0.000010) outperformed existing models, demonstrating MHSA-FN as a scalable, high-performance AI solution for precision agriculture in resource-constrained environments

1. Introduction

Agricultural sustainability and food security are increasingly threatened by the prevalence of crop diseases, which can cause substantial yield losses if not detected early. Hossain et al. [1] illustrated that traditional methods require extensive laboratory analysis, leading to delays in diagnosis. Moreover, the study points out that these techniques often suffer from variability in accuracy due to environmental and human factors, further limiting their reliability. However, deep learning (DL) models require extensive training data and face challenges such as overfitting and high computational costs. Abdu et al.[2] demonstrated that training deep networks' computational demands can be prohibitive, especially in resource-constrained settings. Transfer learning (TL) addresses these limitations by leveraging pre-trained models, reducing computational costs, and enhancing efficiency. Unlike static methods, dynamic models are crucial in adapting to evolving data and emerging disease types. Nguyen et al. [3] illustrated how adaptive learning techniques enable models to update their parameters in response to new disease patterns, improving predictive accuracy over time. However, the limitation of CNNs lies in their inability to capture long-range dependencies. Shah et al. [4] demonstrated how the reliance on limited receptive fields restricts the ability to understand spatial relationships in complex images. This constraint affects tasks requiring a broader contextual understanding, such as medical image analysis. In contrast, the effectiveness of Vision Transformers (ViTs) in modeling global context through self-attention mechanisms was demonstrated by Barman et al. [5]. However, the study also emphasized that ViTs are computationally intensive and require large

datasets for effective training, posing challenges in resource-constrained environments.

Despite these advancements, most current approaches rely on a single CNN model, which may not fully capture local and global feature representations necessary for robust disease detection. Relying solely on CNNs limits their ability to generalize across varying environmental conditions, diverse plant species, and disease variations. Furthermore, He et al. [8] illustrate that CNN-based models suffer from feature redundancy, often struggling to dynamically emphasize critical information in complex images. The study demonstrates that conventional CNN architectures tend to capture redundant or less informative features, reducing model efficiency and hindering performance in tasks requiring fine-grained feature differentiation.

To address these gaps, we propose a framework integrating MobileNetV2, EfficientNetV2, and ViT for parallel feature extraction. MobileNetV2 and EfficientNetV2 extract local and scalable features, respectively, while ViT captures global dependencies directly from the input image. The CNN-extracted features are refined using SE blocks before concatenation with ViT features. A gated control mechanism then regulates the flow of fused features, which are further processed through a multi-scale fusion module (MSFM) to enhance representation learning. Finally, the output is passed through a classification layer for accurate crop disease identification. The primary contributions of this research include:

A tri-branch architecture framework in which MobileNetV2, EfficientNetV2, and ViT operate in parallel to extract both local and global features. Squeeze-and-excitation (SE) blocks enhance channel-wise recalibration, improving feature discrimination. A gated control mechanism and multi-scale fusion module (MSFM**)** that regulate and integrate fused features, an optimized end-to-end learning framework that seamlessly fuses CNN and transformer-based features, enhancing adaptability and scalability, and a Comprehensive statistical validation to confirm the model effectiveness.

The paper is structured as follows: Section 2 discusses related work on CNN-based models, ViTs, and attention-enhanced techniques. Section 3 includes the proposed architecture framework and feature extraction process. Section 4 discusses results analysis, performance evaluations, statistical analysis, and discussion. Section 5 presents the conclusions and outlines directions for future research.

2. Related Work

Ensemble approaches, such as integrating ResNet-18 with multi-head attention, have demonstrated the potential of combining models for enhanced performance. Ramesh et al. [6] utilized a DNN combined with the Jaya Optimization Algorithm, achieving 98.9% classification accuracy for paddy diseases. However, their approach primarily focused on a single crop type, limiting its generalizability across diverse agricultural conditions. Findings by Le et al. [7] indicate that a hybrid convolutional model incorporating liquid neural networks and Neural Circuits achieved 97.15% accuracy while mitigating overfitting. However, the model's complexity may hinder real-time deployment and scalability. A similar study by Guo et al.[8] explored Convolutional Neural Network - Bidirectional Long Short-Term Memory (CNN-BiLSTM) architectures for attention prediction in real-time scenarios, demonstrating that BiLSTMs effectively captured complex dependencies in silica powder moving out of the warehouse. However, these approaches often require extensive computational resources, limiting their deployment on mobile or edge devices. Given the constraints of agricultural environments, lightweight CNN architectures have been explored for real-time crop disease classification. For instance, Bi et al. [9] proposed MobileNetV3 as an efficient model for crop disease detection on edge devices. Amin et al. [10] fused DenseNet121 and EfficientNetB0 through feature concatenation, balancing efficiency and accuracy. Zaki et al. [11] utilized MobileNetV2 to detect various tomato plant diseases by training the model on a complete dataset of images. However, the model performance was limited by diversity and the dataset's quality. Mousavi et al.[12] employed MobileNetV2 for grapevine disease detection, achieving favourable results compared to traditional models like VGG16 and ResNet.

EfficientNetV2 has been applied in various scenarios, particularly for deployment on edge devices like smartphones and drones, which are crucial for real-time applications. Dai et al. [13] implemented EfficientNetV2 for disease detection on drones in precision agriculture. While promising, the approach faced challenges related to model size and real-time processing, which limited its application in field settings. The success of EfficientNetV2, as discussed by Saleem et al. [14], is largely due to its efficient balance between computational demands and high accuracy in agricultural settings. Compared to traditional models, EfficientNetV2, as demonstrated by He et al. [15], requires fewer parameters. This adaptation

ensures improved deployment on resource-constrained devices like embedded systems and smartphones. However, its performance may fluctuate under different environmental conditions, posing challenges to its reliability.

ViTs have shown promise in crop disease detection, surpassing traditional CNNs by overcoming the limitation of local receptive fields. Wang et al. [16] demonstrated how transformers excel in learning long-range dependencies across the entire image. However, despite their enhanced feature extraction capabilities, the computational demands of ViTs and the need for large training datasets pose challenges, particularly in resource-constrained environments. Zhu et al. [17] examined the application of transformers in medical image analysis, focusing on tasks such as classification, segmentation, and detection. Christakakis et al. [18] conducted a detailed study using ViT-based deep learning models to detect Botrytis cinerea in Cucurbitaceae crops, attaining an accuracy of 92%. However, the variance performance of the model may vary with diverse crop types and field conditions, limiting its broader applicability in diverse agricultural settings. Barman et al. [5] implemented an edge-based disease detection approach using a ViT model trained on 10,010 tomato leaf image datasets. The study compared the ViT and Inception V3 models for classifying 10 disease types, with the ViT model achieving 90.99% accuracy, outperforming Inception V3. While the integration into an Android app demonstrated the model's potential for large-scale agricultural applications, challenges remained in ensuring robustness across varying environmental conditions and optimizing the model for real-time performance.

Wang et al. [19] argue that gated mechanisms play a key role in DL models by enabling them to focus on key features while suppressing inappropriate information. This leads to improved model performance and interpretability. However, the added complexity of these mechanisms may require additional computational resources and fine-tuning, potentially impacting the model's efficiency and scalability in resource-constrained environments. Wang et al. [20] explored the use of gated mechanisms, specifically Gated Recurrent Unit (GRU), for classifying plant diseases based on leaf images at different growth stages. While the method improved feature selection by focusing on relevant patterns, it faced challenges in handling large-scale datasets and generalizing across various environmental conditions. Bi et al. [21] applied a combination of gated convolutional layers and recurrent units for detecting diseases in soybean crops. This hybrid approach effectively captured spatial and temporal features, but its computational demands may limit its scalability for real-time applications in field conditions, especially with large-scale datasets.

Attention mechanisms have been crucial for crop disease detection in DL, enabling model improvement in efficiency and accuracy. Ni et al. [22] proposed a MaizeHT model by combining CNN with self-attention for accurate maize growth stage recognition. The model utilized ResNet34 for feature extraction and applied multi-head self-attention to predict growth stages. MaizeHT achieved 97.71% accuracy with 224×224 resolution and 98.71% accuracy with 512×512 resolution on a self-built dataset and with 15.446 million parameters. Wang et al. [23] introduced a crop mapping model that incorporated temporal and spatial modules for feature extraction, along with multi-head self-attention and positional encoding. The model demonstrated improved accuracy, surpassing Transformers by 3.35%, LSTM models by 6.42%, and CNNs by 1.40%. However, the increased model complexity and reliance on attention mechanisms may result in higher computational demands, limiting its scalability for large-scale, real-time applications.

This paper addresses these critical gaps in existing plant disease detection models, enhancing their adaptability, efficiency, and accuracy. Firstly, while prior works such as Jouini et al. [24] and Hassan et al. [25] have explored CNN-based or transformer-based architectures individually, they often lack an optimal fusion strategy that fully leverages the strengths of both approaches. This limitation results in suboptimal feature extraction, mainly when dealing with complex plant disease patterns requiring local texture details and global spatial relationships. Secondly, existing models frequently struggle with effective feature recalibration, leading to inefficient utilization of learned representations. While attention mechanisms, as in the works of Ansari et al. [26], like SE blocks and CBAM, have been used, they are often applied in isolation rather than as part of a structured multi-level attention framework that prioritizes key features across multiple dimensions. Thirdly, Wang et al. [27] correctly argue that many current approaches lack robust multi-scale feature fusion, limiting their ability to handle variations in disease severity, lighting conditions, and background noise in real-world images. While inception-based architectures, such as in the works of Shah et al. [4], have shown promise in general image processing tasks, their application in plant disease detection remains underexplored.

3. Methods

3.1 Proposed Model Architecture

The hybrid Tri-branch multimodal design, developed for enhanced feature extraction, is illustrated in Figure 1. This architecture integrates two parallel convolutional backbones, MobileNetV2 and EfficientNetV2, which focus on extracting localized and hierarchical features. These models work in tandem to extract a broad spectrum of features, which are then refined through an SE block that reweights channels, amplifying the most relevant ones and reducing the impact of irrelevant features. A ViT block processes image patches parallel with convolutional backbones to capture global and long-range dependencies across the entire image. The local features extracted are then concatenated. This allows the model to leverage detailed local information and broader global patterns.

An attention gate mechanism is applied to focus further the model's attention on key areas of the image. This mechanism learns attention maps to show the most important areas of the image, effectively suppressing background noise and irrelevant features. Following this, a multi-scale fusion module aggregates fine-grained, medium, and coarse features using multiple receptive fields, ensuring that the model can capture information at various levels of abstraction.

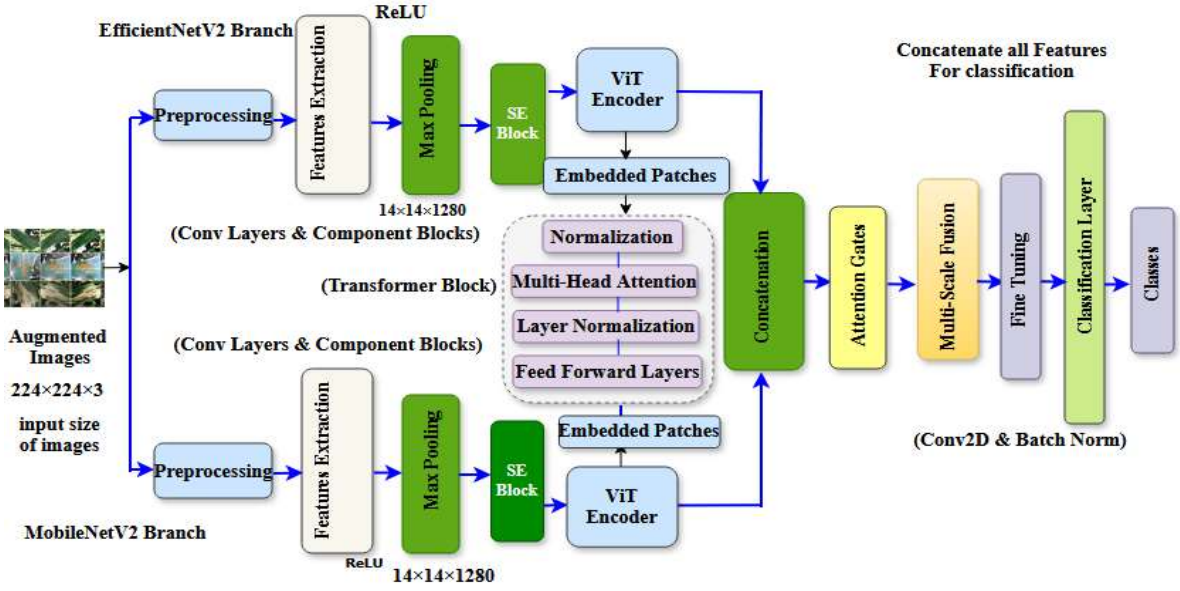

Figure 1. Proposed Model Architecture.

3.2. Data Preparation and Preprocessing

The data was loaded from a directory structure where images were organized in subfolders based on their class labels, following a standard format. Each subfolder represented a distinct disease category, ensuring efficient label assignment during model training. This hierarchical organization streamlined data retrieval facilitated preprocessing, and enabled seamless integration with deep learning pipelines. TensorFlow's image dataset directory function was used to load the data with shuffling, batching, and resizing. As shown in Table 1, off-the-fly data augmentation was applied to enhance model generalization during training. This included random rotation, flipping, brightness adjustment, and zoom transformations, introducing variability to simulate real-world conditions and address class imbalance.

The images were resized to a standard input size of 224×224 pixels and normalized to the [0,1] range by dividing pixel values by 255.0. This preprocessing step ensured uniformity across all input images. A stratified sampling approach was applied to maintain a consistent class distribution within the training and validation datasets. This helped prevent biases that could arise from class imbalances. The number of samples allocated for the training set was determined using stratified sampling, ensuring proportional representation across dataset splits. Three backbone models were utilized for feature extraction, as shown in Figure 1: MobileNetV2, EfficientNetV2, and ViT. MobileNetV2 and EfficientNetV2, both lightweight and efficient architectures, were pre-trained on ImageNet, while ViT processed images as sequences of patches using self-attention mechanisms. A Squeeze-and-Excitation (SE) block was incorporated to enhance feature representation, recalibrating feature maps by capturing channel dependencies. This involved computing channel-wise statistics through global average pooling, followed by two fully connected layers with ReLU activation and a sigmoid function to generate

attention weights. The attention vector was then applied to the input feature map, refining its representation.

A multi-fusion module was integrated, leveraging convolutional operations with different kernel sizes (1×1, 3×3, and 5×5), along with max pooling, to capture multi-scale features. The extracted features from MobileNetV2, EfficientNetV2, and ViT were concatenated to form a comprehensive feature representation. The merged features were then processed through a fully connected layer, batch normalization, and dropout to mitigate overfitting. The final classification layer employed the softmax function to predict class probabilities, ensuring that outputs corresponded to the number of predefined classes. The AdamW optimizer was used with a fine-tuned learning rate and weight decay for model optimization. The model was trained using categorical cross-entropy as the loss function, which effectively handled multi-class classification tasks by comparing predicted probabilities against actual labels. Several training techniques were employed to enhance model performance and stability. ModelCheckpoint was used to save the best model based on validation loss. The model with the lowest validation loss was preserved during training, ensuring that the most optimal parameters were retained. EarlyStopping was implemented to halt training if the validation loss did not improve for a predefined number of epochs. This helped prevent overfitting and conserved computational resources by stopping training once performance plateaued. ReduceLROnPlateau dynamically adjusted the learning rate when validation loss stopped improving. If the loss remained stagnant for a set number of epochs, the learning rate was reduced by a predefined factor. This adjustment allowed the model to refine its parameters more effectively, leading to better convergence and improved overall performance.

3.3 Dataset Description

This study combined the Kaggle dataset [28] with a locally developed dataset, FieldPlant, to create a comprehensive resource named the DEMF dataset. The FieldPlant dataset consisted of 25,775 images, as detailed in Table 1, while the Kaggle dataset, comprising 38 distinct classes and 60,343 images (see Table 2), provided a globally diverse and well-structured resource ideal for training and validating the models. Annotated images of plant leaves were collected over different seasons from farms in Central Kenya. April–May and October–November was prioritized for capturing fungal and bacterial diseases, while June–July and December–January focused on chronic infections like viral diseases and stress-related symptoms. A plant pathologist classified the images into different disease categories, annotated them, and organized them into folders. A combination of agricultural, ecological, and methodological factors drove the decision to collect field images from six counties in central Kenya. These counties span diverse agroecological zones, representing varying climatic conditions, soil types, and farming practices.

Table 1. Field Plant Dataset.

Crop	Disease	Count	Region
Banana	Cordana	483	Murang'a
Banana	Healthy	479	Murang'a
Banana	Pestalotiopsis	480	Murang'a
Banana	Sigatoka	507	Murang'a
Bean	Angula-Leaf	377	Kiambu
Bean	Rust	403	Kiambu
Bean	Healthy	408	Kiambu
Cassava	Brown Spot	1533	Murang'a
Cassava	Green Mite	1152	Murang'a
Cassava	Healthy	1415	Murang'a
Cassava	Mosaic	1305	Murang'a
Maize	Grasshopper	707	Kirinyaga
Maize	Fall Army	331	Kirinyaga
Maize	Healthy	271	Kirinyaga
Maize	Leaf Beetle	1181	Kirinyaga
Maize	LeafBlight	1151	Kirinyaga
Maize	Streak Virus	1164	Kirinyaga
Maize	Leaf Spot	1453	Kirinyaga
Rice	LeafBlight	114	Kirinyaga
Rice	Brown Spot	118	Kirinyaga
Rice	Healthy	123	Kirinyaga
Rice	LeafBlast	112	Kirinyaga
Rice	Brown Spot	116	Kirinyaga
Sugarcane	Healthy	614	Embu
Sugarcane	Mosaic	534	Embu
Sugarcane	Red Rot	606	Embu
Sugarcane	Rust	606	Embu
Sugarcane	Yellow	591	Embu
Tea	Algal Leaf	138	Nyeri
Tea	Anthracnose	114	Nyeri
Tea	Eye Spot	113	Nyeri
Tea	Bro- blight	132	Nyeri
Tea	Healthy	79	Nyeri
Tea	Leaf Spot	169	Nyeri
Sunflower	Downy mild	139	Meru
Sunflower	Fresh Leaf	147	Meru
Sunflower	Gray Mold	80	Meru
Sunflower	Leaf Scars	160	Meru
Maize	Streak Virus	1154	Muranga
Maize	Grasshopper	794	Muranga
Maize	Fall-Army	336	Muranga
Cassava	Green Mite	1196	Muranga
Cassava	Brown Spot	1746	Muranga
Banana	Cordana	472	Kirinyaga
Banana	Sigatoka	472	Kirinyaga
Total		**25775**	

The Kaggle dataset is highly valuable for crop disease detection due to its extensive variety, covering multiple crops and diseases with a well-balanced representation of healthy and diseased samples. With 60,343 labeled instances, it provides a robust foundation for training machine learning models, ensuring diversity across plant species and disease types. The dataset's class distribution,

particularly for critical diseases like Huanglongbing in oranges and Tomato Leaf Curl Virus, supports effective model generalization. Additionally, the inclusion of healthy samples across crops aids in distinguishing between diseased and non-diseased conditions, enhancing real-world applicability for precision agriculture solutions.

Table 2. Kaggle Dataset.

Crop	Disease	Count
Apple	Apple Scab	1000
Apple	Black Rot	1000
Apple	Cedar Apple Rust	1000
Apple	Healthy	1645
Blueberry	Healthy	1502
Cherry	Powdery Mildew	1052
Cherry	Healthy	1000
Corn	Cercospora Spot	1000
Corn	Common Rust	1192
Corn	Northern Blight	1000
Corn	Healthy	1162
Grape	Black Rot	1180
Grape	Esca Black Measles	1383
Grape	Leaf Blight	1076
Grape	Healthy	1000
Orange	Huanglongbing	5507
Peach	Bacterial Spot	2297
Peach	Healthy	1000
Pepper	Bacterial Spot	1478
Pepper	Healthy	1000
Potato	Early Blight	1000
Potato	Late Blight	1000
Potato	Healthy	1000
Raspberry	Healthy	1000
Soybean	Healthy	5090
Squash	Powdery Mildew	1835
Strawberry	Leaf Scorch	1109
Strawberry	Healthy	1000
Tomato	Bacterial Spot	2127
Tomato	Early Blight	1000
Tomato	Late Blight	1909
Tomato	Leaf Mold	1000
Tomato	Septoria Leaf Spot	1771
Tomato	Spider Mites	1676
Tomato	Target Spot	1404
Tomato	Tomato Leaf Curl	5357
Tomato	Tomato Mosaic Virus	1000
Tomato	Healthy	1591
Grand Total		**60343**

To create the final combined dataset, augmentations were applied to classes with fewer images. Rotation involves rotating the images by a specified angle (e.g., 0–40 degrees) to help the model learn rotational invariance. Shear distorts the image along the x or y axis by a certain angle to simulate 3D transformations, with a proposed 0-20 degrees value. Zooming adjusts the image scale by a factor of 0.8-1.2, allowing the model to handle different object sizes. Horizontal Flip mirrors the image to enhance recognition of objects in reversed orientations, with a proposed value of True. Vertical Flip functions similarly but is less commonly used, with a proposed value of False. The Final dataset, as shown in Table 3, consisted of 99,551 images of 22 different crop types, including apples, bananas, beans, corn, maize, grapes, and tomatoes, with the largest category being tomatoes at 18,841 images.

Table 3. Combined Dataset.

Crop Type	Total Images	Training Images	Validation Images
Apple	4,651	3,719	932
Banana	4,008	3,204	804
Beans	8,096	6,475	1,621
Blueberry	1,502	1,201	301
Cassava	4,894	3,914	980
Cherry	2,054	1,642	412
Corn	4,358	3,484	874
Grape	4,641	3,711	930
Maize	1,002	801	201
Maize-L	1,239	991	248
Maize	4,985	3,986	999
Orange	5,507	4,405	1,102
Peach	3,299	2,638	661
Pepper	2,480	1,983	497
Potatoes	3,006	2,403	603
Raspberry	1,002	801	201
Rice	5,010	4,005	1,005
Squash	1,835	1,468	367
Strawberry	2,111	1,688	423
Sugarcane	5,010	4,005	1,005
Sunflower	4,008	3,204	804
Tea	6,012	4,806	1,206
Tomatoes	18,841	15,067	3,774
Total	**99,551**	**79,601**	**19,950**

3.4 Experimental Parameters and Environment

Table 4 presents the hyperparameter configurations used in our experiments. The model was trained on images resized to 224 × 224 with three channels to maintain consistency across architectures. The ViT encoder was designed with six layers, each containing eight multi-head self-attention blocks. A patch size of 7 was used to segment feature maps, with an embedding dimension of 128 to balance computational efficiency and feature representation. The multi-layer perceptron (MLP) dimension was set to 2048 and 0.5 dropout rate. The model was trained for 17 epochs using the AdamW optimizer with an initial learning rate of 0.00001 and a weight decay 0.0001. Experiments were conducted on an NVIDIA RTX 3090 GPU, ensuring efficient performance. Additionally, the setup included a virtualized Intel Xeon CPU with access to virtualized GPUs, including NVIDIA T4, Tesla P100, and K80.

Table 4. Hyperparameter Configurations.

Hyperparameter	Value
Image size	224 × 224
Image channels	3
Patch size	7
ViT encoder layers	6
Number of multi-head self-attention blocks	8
Hidden dimensions	128
Layer perceptron dimension	2048
Dropout rate	0.5
Epochs	17

3.5. Evaluation Approach

A diverse range of metrics was selected to assess the model's predictive accuracy and overall

performance. Accuracy measured the proportion of correctly classified instances, providing an overall effectiveness indicator. Precision evaluated the proportion of correctly predicted positive cases out of all positive predictions, ensuring the model's reliability in identifying diseases. Recall assessed the percentage of positive cases detected, which is crucial in disease detection to minimize false negatives and prevent missed diagnoses. The F1 score balanced precision and recall, making it particularly valuable when false positives and negatives carry significant implications. The ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) quantified how well the model differentiated between categories, with a higher AUC score (closer to 1.0) indicating superior classification accuracy.

4. Results and Discussion

4.1. Classification Results

Table 5 presents the model's training and validation results, demonstrating notable performance improvements. These findings emphasize the model robustness and efficiency during training. The model demonstrated steady improvement over 17 epochs, consistently increasing training and validation performance. Training loss decreased from 2.1238 in the first epoch to 0.8382 in the final epoch, while validation loss also shows a downward trend, indicating effective optimization. Accuracy improved significantly, with training accuracy rising from 61.31% to 99.61% and validation accuracy increasing from 89.56% to 98.66%, suggesting strong generalization. The learning rate remained constant at 1e-5, ensuring stable training without sudden fluctuations.

Table 5. Training and Validation Performance.

Epoch	Loss	Accuracy (%)	Val.Loss	Val.Acc (%)	L.Rate
1	2.1238	61.31	1.1773	89.56	1e-5
2	1.2297	88.15	1.0571	94.89	1e-5
3	1.0932	93.02	0.9983	96.37	1e-5
4	1.0306	95.06	0.9615	96.95	1e-5
5	0.9886	96.30	0.9462	97.54	1e-5
6	0.9581	97.18	0.9218	97.94	1e-5
7	0.9347	97.85	0.9061	98.06	1e-5
8	0.9171	98.23	0.8912	98.27	1e-5
9	0.9030	98.51	0.8836	98.31	1e-5
10	0.8899	98.76	0.8763	98.37	1e-5
11	0.8783	99.03	0.8697	98.37	1e-5
12	0.8711	99.15	0.8648	98.45	1e-5
13	0.8626	99.27	0.8556	98.56	1e-5
14	0.8559	99.33	0.8503	98.52	1e-5
15	0.8496	99.42	0.8472	98.55	1e-5
16	0.8443	99.43	0.8453	98.55	1e-5
17	0.8382	99.61	0.8368	98.66	1e-5

The training and validation loss graphs, as shown in Figure 2 and Figure 3, typically showed a steady decline, indicating that the model was effectively learning from the data.

Figure 2. Training and Validation Loss.

Figure 3. Training and Validation Loss.

Table 6(a). Classification Performance for Classes (1-18).

Class Name	Precision	Recall	F1-Score	Support
Apple_Apple-scab	1.00	1.00	1.00	200
Apple_Black-rot	1.00	1.00	1.00	200
Apple_ apple-rust	1.00	1.00	1.00	200
Apple___healthy	1.00	1.00	1.00	329
Banana_cordana	1.00	0.98	1.00	200
Banana-healthy	0.99	1.00	1.00	200
Banana_pestalotiopsi	0.99	0.99	0.99	200
Banana_sigatoka	1.00	0.99	1.00	200
Bean_angular spot	1.00	0.97	0.98	201
Beans_healthy	1.00	0.99	1.00	200
Blueberry_healthy	1.00	1.00	1.00	301
Cassava_brown	1.00	1.00	1.00	296
Cassava_mite	0.99	0.95	0.96	203
Cassava healthy	0.98	0.99	0.98	239
Cassava -mosaic	0.96	0.98	0.97	241
Cherry_Powdery_	1.00	1.00	1.00	211
Cherry_healthy	1.00	0.99	1.00	200
Corn_healthy	1.00	0.99	0.99	233

Table 7(b). Classification Performance for Classes (19-40).

Class Name	Precision	Recall	F1-Score	Support
Corn_leaf_spot	0.97	0.96	0.96	201
Corn_Common_rust	1.00	1.00	1.00	239
Corn _Leaf_Blight	0.95	0.97	0.96	200
Corn_healthy	1.00	0.99	0.99	233
Grape Black_rot	0.99	1.00	1.00	236
Grape Esca	1.00	0.99	1.00	277
Grape Leaf_blight	1.00	1.00	1.00	215
Grape healthy	1.00	1.00	1.00	200
Maize _grasshoper	0.99	0.98	0.99	200
Maize_leaf_spot	0.75	0.68	0.72	248
Maize_fall_Army	0.95	1.00	0.97	201
Maize-healthy	0.92	1.00	0.95	199
Maize_leaf_beetle	0.97	0.97	0.97	199
Maize-leaf_bligh	0.74	0.73	0.72	200
Maize_streak-virus	0.91	0.89	0.90	199
Orange_Haunglongb	1.00	1.00	1.00	1102
Peach _spot	1.00	1.00	1.00	460
Peach_healthy	1.00	1.00	1.00	200
Pepper_spot	1.00	0.99	1.00	200
Pepper _healthy	1.00	1.00	1.00	296
Potato_Early_blight	1.00	1.00	1.00	200
Potato_Late_blight	1.00	0.99	0.99	201

Table 7(c). Classification Performance for Classes (41-70).

Class Name	Precision	Recall	F1-Score	Support
Potato_healthy	0.99	1.00	0.99	200
Raspberry healthy	1.00	0.99	1.00	201
Rice-bact blight	1.00	1.00	1.00	200
Rice_brown_spot	0.99	0.98	0.98	201
Rice_healthy	1.00	1.00	1.00	201
Rice_leaf_blast	0.98	0.99	0.99	200
Rice _brown_spot	1.00	1.00	1.00	200
Soybean_healthy	1.00	1.00	1.00	1018
Squash_ mildew	1.00	0.99	1.00	367
Strawberry/scorch	0.99	1.00	1.00	222
Strawberry_healthy	1.00	1.00	1.00	200
Sugarcane_Healthy	0.99	0.99	0.98	200
Sugarcane_Mosaic	0.98	0.98	0.97	201
Sugarcane_RedRot	0.99	1.00	0.99	200
Sugarcane_Rust	1.00	0.96	0.98	200
Tea_Anthracnose	1.00	1.00	1.00	201
Tea_algal-leaf	0.97	0.98	0.96	200
Tea_bird eye spot	0.98	0.99	1.00	201
Tea_brown-blight	1.00	1.00	1.00	200
Tea_healthy	1.00	1.00	1.00	200
Tea_redleaf-spot	1.00	1.00	1.00	200
Tomato _spot	0.99	1.00	1.00	426
Tomato_Early	0.99	0.96	0.97	201
Tomato_Lateblight	0.99	0.99	0.99	382
Tomato-Target_	0.99	0.99	0.99	281
Tomato- Curl_	1.00	1.00	1.00	1072
Tomato_mosaic_	1.00	1.00	1.00	200
Tomato-healthy	1.00	1.00	1.00	318
bean_rust	0.97	1.00	0.98	201

The confusion matrices shown in Figure 4(a-f) illustrate the performance of the disease detection task across all classes (ranging from class 0 to class 70). These matrices display actual class labels on the X-axis and predicted labels on the Y-axis, providing insights into the model's classification accuracy for each class.

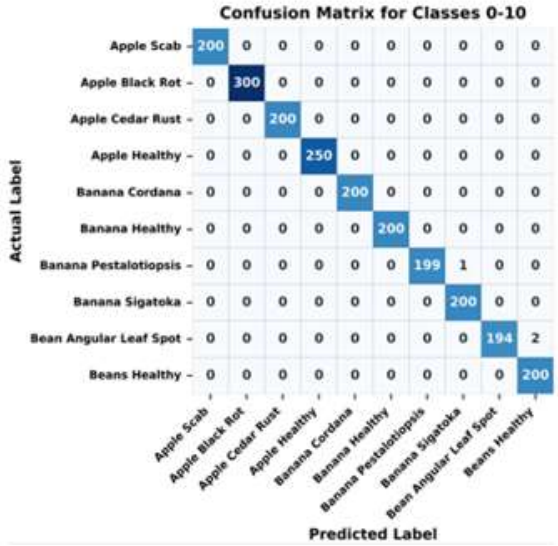

Figure 4(a). Confusion Matric for Classes 0-10.

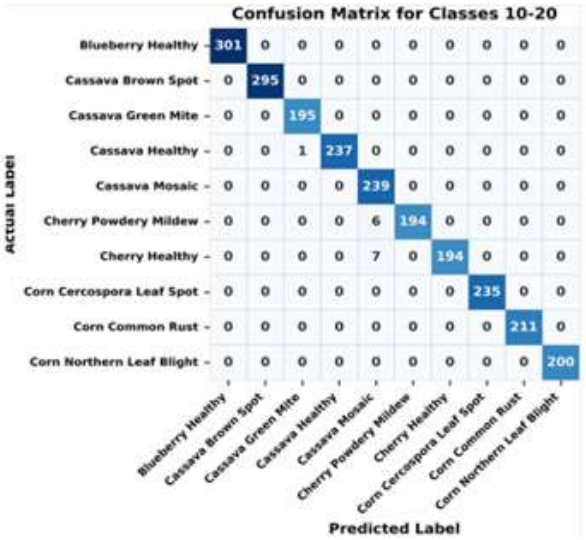

Figure 4(b). Confusion Matric for Classes 10-20.

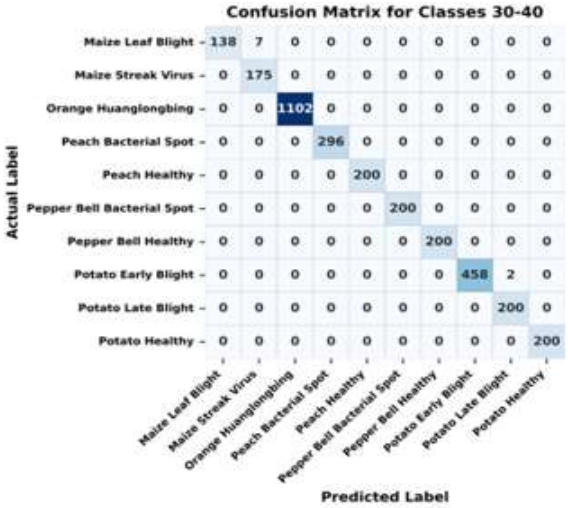

Figure 4(c). Confusion Matric for Classes 30-40.

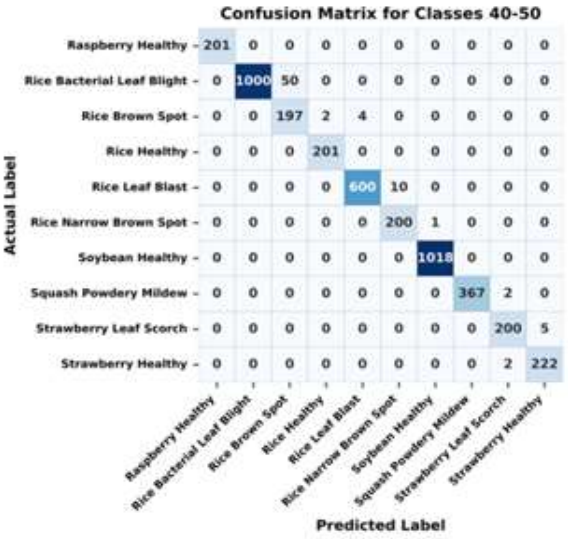

Figure 4(d). Confusion Matric for Classes 40-50.

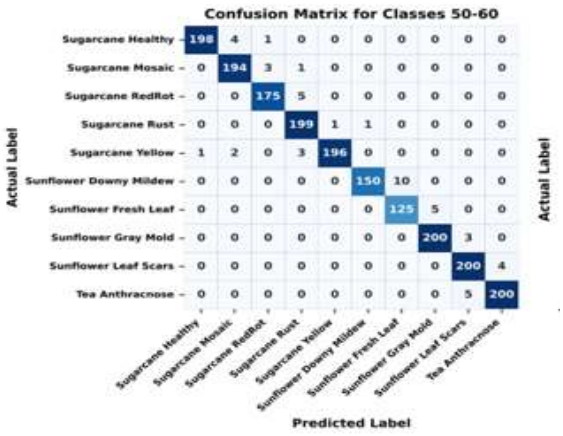

Figure 4(e). Confusion Matric for Classes 50-60.

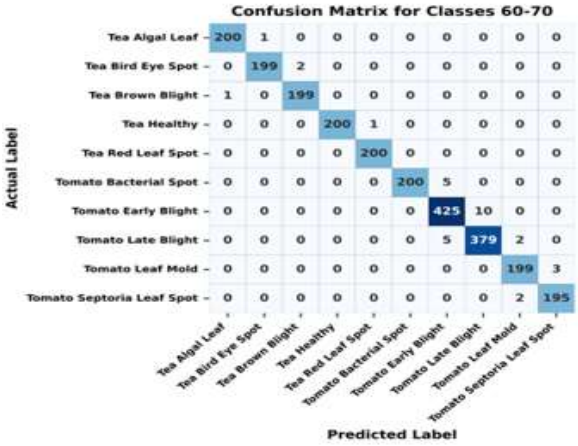

Figure 4(f). Confusion Matric for Classes 60-70.

The AUC score reflects the model's ability to effectively differentiate between positive and negative classes, with higher AUC values indicating stronger predictive power and better overall performance. An AUC score closer to 1.0 suggests superior classification ability, whereas lower values indicate room for improvement. The per-class precision, recall, and F1 score provide a more granular evaluation of the model's strengths and weaknesses.

Figure 5(a). ROC AUC Scores.

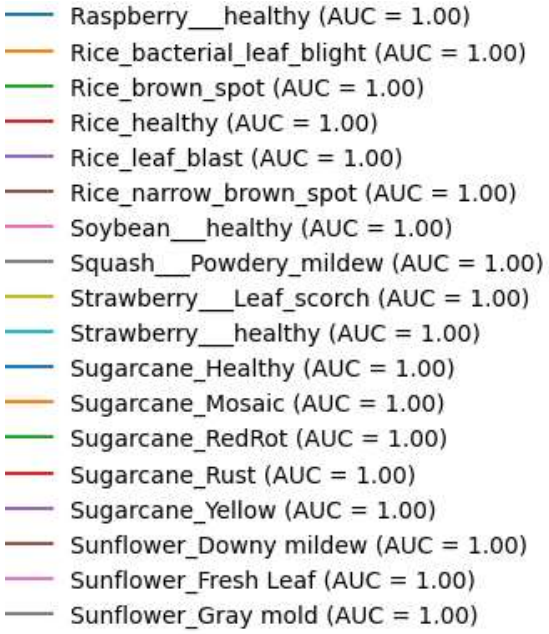

Figure 5(b). ROC AUC Scores.

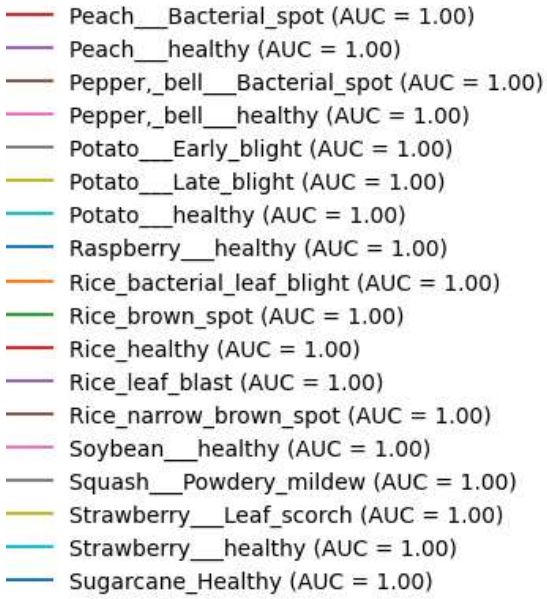

Figure 5(c). ROC AUC Scores.

4.2. Ablation Studies

Ablation studies evaluated the impact of components like the Multi-scale Module, Gated Mechanism, SE blocks, and ViT on model performance. The most effective configurations for crop disease detection were identified by systematically adding or removing these elements. The baseline CNN model with EfficientNetV2 and MobileNetV2 achieved 98.55% accuracy, slightly dropping to 98.45% with minor modifications. Introducing a Multi-scale Module improved accuracy to 98.57% while adding SE attention and a Gated Mechanism further enhanced it to 98.66%, as shown in Table 10. Parameter analysis showed that the baseline model had 8.85M parameters while integrating all enhancements, which

increased complexity to 38.86M. Extraction mechanisms. The exclusion of the Multi-scale Module and Gated Mechanism resulted in parameter reductions to 11.87M and 14.23M**,** respectively, showing their contribution to model size as shown in Table 11.

Table 8. Impact of Different Architectural Modifications.

Model Number	Model Configuration	Multi-scale Module	Gated Mechanism	Accuracy
1	CNN model: EfficientNetV2 and MobileNetV2	No	No	98.55%
2	CNN models: EfficientNetV2 and MobileNetV2 (Epochs Reduction)	No	No	98.45%
3	CNN models: EfficientNetV2 and MobileNetV2, with Multiscale module	Yes	No	98.57%
4	**Proposed Model:** EfficientNetV2 and MobileNetV2, with Multi-scale module, SE, and Gated Mechanism	Yes	Yes	**98.66%**

Table 9. Parameters Comparison and Training Configurations.

Model Configuration	Total Parameters	Trainable Parameters	Non-Trainable Parameters	Epochs	Batch Size	Learning Rate
EfficientNetV2 + MobileNetV2	8,853,468	8,758,236	95,232	17	4978	0.00001
EfficientNetV2 + MobileNetV2 + SE + ViT + Gated Mechanism + Multiscale Module	38,863,998	38,767,742	96,256	17	4978	0.00001
EfficientNetV2 + MobileNetV2 + SE + ViT + Gated Mechanism + Multiscale Module (No Multiscale Module)	11,871,964	11,776,220	95,744	17	4978	0.00001
EfficientNetV2 + MobileNetV2 + SE + ViT + Gated Mechanism (No Gated Mechanism)	14,227,932	14,132,700	95,232	17	4978	0.00001

Data augmentation techniques were also carried out to test the influence on the model's performance. Figures 6 and 7 show that each transformation aimed to introduce variability, simulating real-world conditions. The rotation randomly altered the leaf's orientation, while flipping helped the model generalize across orientations. Brightness adjustment simulated lighting conditions, and zoom introduced scale and focus variation.

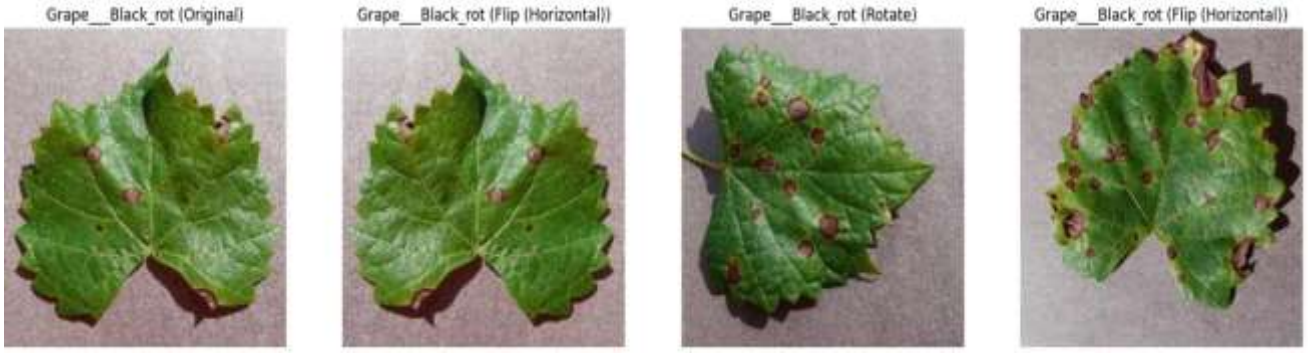

Figure 6. Grape Leaf Sample Augmentation.

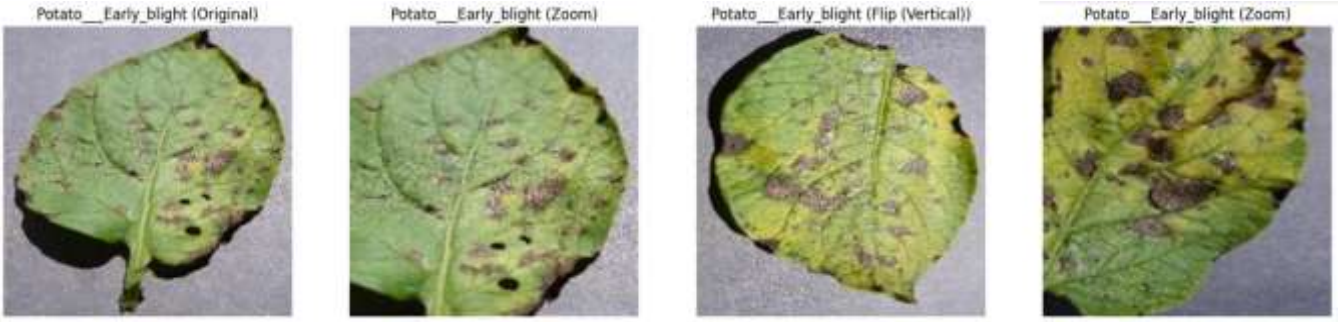

Figure 7. Potato Leaf Augmentation.

On the unseen data, 249 images were processed, with 232 correctly classified, resulting in an accuracy of 93.17%. Only 17 images (6.83%) were misclassified. Further supporting the model's strong performance overall. The model demonstrated its ability to accurately classify plant diseases, even when the confidence scores were low for a few classes lacking dominant features, as shown in Figure 8. This suggested that the proposed model, as shown in Figures 9 and 10, generally distinguished between healthy and diseased plants from the unseen data.

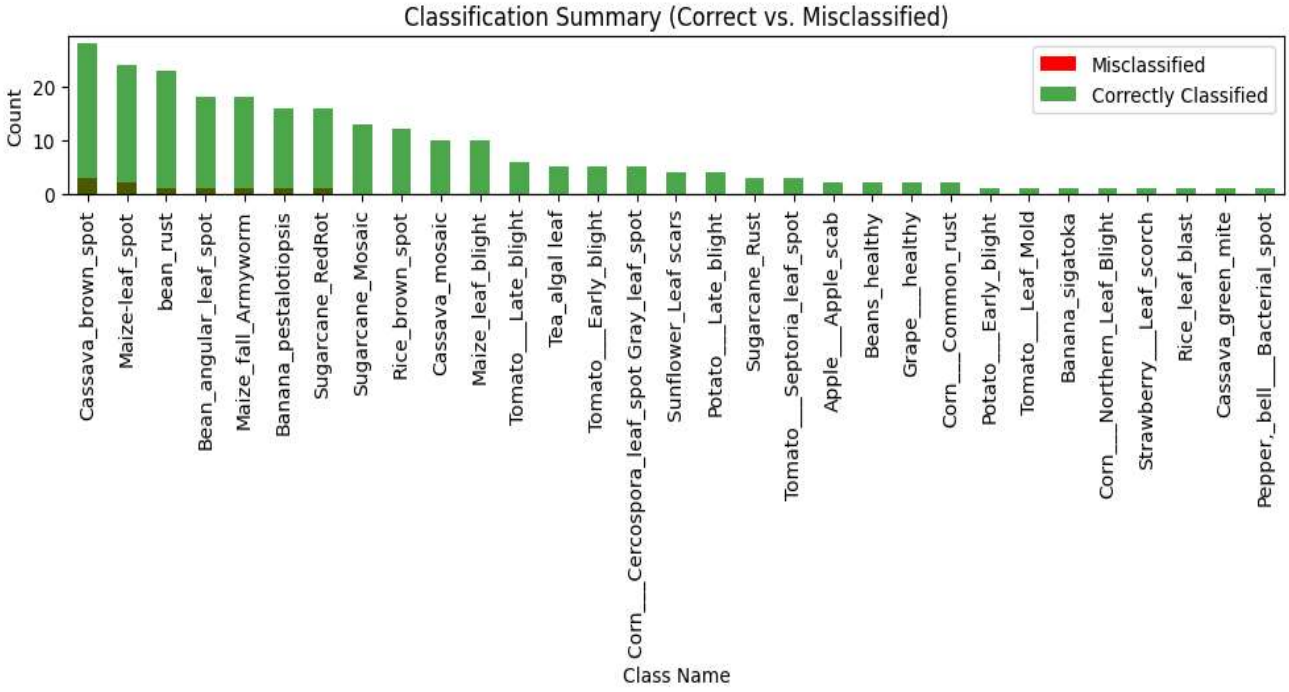

Figure 8. Test Classification Summary.

Figure 9. Mixed Random Classes Actual Vs. Predicted Classification.

Figure 10. Beans Class Actual Vs. Predicted Classification.

4.3. Statistical Testing

Statistical testing evaluated the significance of performance differences across model variations. The performance metrics, as shown in Table 10, demonstrate the superiority of the proposed model across all evaluation parameters. The evaluation metrics, including accuracy, precision, recall, F1-score, Cohen's Kappa, and AUC, provided a comprehensive assessment of model performance. The proposed model outperformed other architectures across all metrics, with an 85.1% Bayesian superiority probability over ShuffleNet, which was second, reinforcing its selection as the optimal model for deployment.

Table 10. Statistical Comparison of the Models.

Model	Accuracy	Precision	Recall	F1-score	Kappa	AUC	Rank
Proposed Model	**0.990**	**0.993421**	**0.990085**	**0.990365**	**0.989855**	**0.999997**	**1st**
ShuffleNet	0.982	0.983906	0.982153	0.980596	0.981740	0.999991	2nd
EfficientNetV2	0.972	0.976770	0.973841	0.973155	0.971595	0.999935	3rd
VGG-16	0.972	0.976770	0.973841	0.973155	0.971595	0.999935	3rd
DenseNet	0.956	0.966270	0.962312	0.958693	0.955368	0.999863	4th
AlexNet	0.942	0.953236	0.947609	0.942789	0.941164	0.998949	5th
DenseNet50	0.884	0.907292	0.889661	0.883751	0.882337	0.998823	6th

4.4. Statistical Analysis of Confidence Scores

Confidence variance reflects the consistency of a model's classification confidence. Lower variance indicates greater stability in predictions, reducing fluctuations in confidence levels. The results show The proposed model had the lowest confidence variance, as shown in Table 11 and Figure 11.

Table 11. Confidence Variance Across Models.

Model	Confidence Variance
DenseNet50	0.000034
AlexNet	0.000030
DenseNet	0.000023
EfficientNetV2	0.000016
VGG_16	0.000016
ShuffleNet	0.000013
Proposed Model	**0.000010**

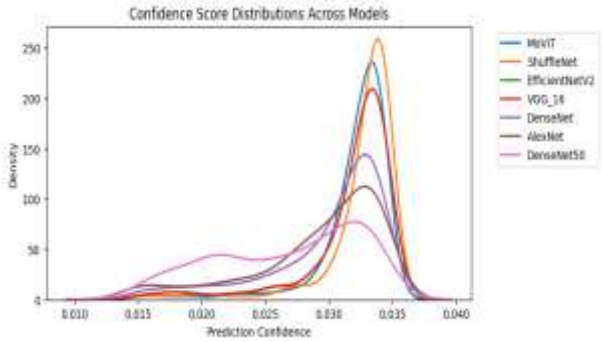

Figure 11. Confidence score Distribution.

4.5. Comparison with Existing Models

The proposed model consistently outperformed existing crop disease detection and related models we used in our study, as summarized in Table 12 and Table 13. It demonstrated superior accuracy, efficiency, and robustness across multiple evaluation metrics. These results highlight its effectiveness in real-time crop disease detection, making it a reliable solution for agricultural applications.

Table 12. Comparison with Existing Models.

Model	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
Proposed Model	**99.61%**	**98.66%**	**0.0331**	**0.8458**
MobileNetV2	98.92%	98.21%	0.0507	0.8663
EfficientNetB0	97.65%	91.02%	0.0811	1.2424
EfficientNetV2	99.08%	97.95%	0.0702	1.0291
DenseNet121	98.80%	97.75%	0.0675	1.0733
DenseNet50	98.75%	96.11%	0.0706	1.0975
ResNet152	98.74%	96.45%	0.0852	1.2092
AlexNet	97.88%	93.50%	0.1189	1.5391
Custom CNN	92.10%	61.84%	0.2750	2.5675

Table 13. Comparison with State-of-the-Art Models.

Related Studies	Classification Accuracy
[28]	98.00%%
[17]	97.50%%
[29]	90.00%
[30]	90.99%
[31]	85.02%
The proposed model	**98.66%**

5. Conclusion

This study demonstrated the potential of hybrid deep learning models integrating MobileNetV2, EfficientNetV2, and Vision Transformers (ViT) for state-of-the-art crop disease classification. The proposed model achieved 99.0% accuracy and an AUC of 0.999997, excelling in distinguishing healthy from diseased crops and combining lightweight CNNs with ViT's self-attention balanced speed and accuracy, outperforming ShuffleNet and VGG-16 while maintaining low computational costs. Squeeze-and-excitation blocks and gated attention further improved feature focus, enabling detection in complex field

conditions. However, challenges remain in generalizing across diverse environments, particularly underrepresented disease strains and real-world complexities. Future research should explore self-supervised learning to reduce reliance on annotated datasets, dynamic inference for resource-efficient deployment, and expanded datasets through agricultural collaborations. Addressing these issues will enhance real-world reliability, support scalable, AI-driven crop health monitoring, and promote global food security.

References

[1] M. I. Hossain, S. Jahan, M. R. Al Asif, M. Samsuddoha, and K. Ahmed, "Detecting tomato leaf diseases by image processing through deep convolutional neural networks," *Smart Agricultural Technology*, vol. 5, Oct. 2023, doi: 10.1016/j.atech.2023.100301.

[2] A. M. Abdu, M. M. Mokji, and U. U. Sheikh, "Machine learning for plant disease detection: An investigative comparison between support vector machine and deep learning," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 4, pp. 670–683, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp670-683.

[3] H. T. Nguyen, H. H. Luong, L. B. Huynh, B. Q. H. Le, N. H. Doan, and D. T. D. Le, "An Improved MobileNet for Disease Detection on Tomato Leaves," *Advances in Technology Innovation*, vol. 8, no. 3, pp. 192–209, 2023, doi: 10.46604/aiti.2023.11568.

[4] S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, and F. Marinello, "Comparing Inception V3, VGG 16, VGG 19, CNN, and ResNet 50: A Case Study on Early Detection of a Rice Disease," *Agronomy*, vol. 13, no. 6, Jun. 2023, doi: 10.3390/agronomy13061633.

[5] U. Barman *et al.*, "ViT-SmartAgri: Vision Transformer and Smartphone-Based Plant Disease Detection for Smart Agriculture," *Agronomy*, vol. 14, no. 2, Feb. 2024, doi: 10.3390/agronomy14020327.

[6] S. Ramesh and D. Vydeki, "Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm," *Information Processing in Agriculture*, vol. 7, no. 2, pp. 249–260, Jun. 2020, doi: 10.1016/j.inpa.2019.09.002.

[7] A. T. Le, M. Shakiba, I. Ardekani, and W. H. Abdulla, "Optimizing Plant Disease Classification with Hybrid Convolutional Neural Network–Recurrent Neural Network and Liquid Time-Constant Network," *Applied Sciences (Switzerland)*, vol. 14, no. 19, Oct. 2024, doi: 10.3390/app14199118.

[8] D. Guo, P. Duan, Z. Yang, X. Zhang, and Y. Su, "Convolutional Neural Network and Bidirectional Long Short-Term Memory (CNN-BiLSTM)-Attention-Based Prediction of the Amount of Silica Powder Moving in and out of a Warehouse," *Energies (Basel)*, vol. 17, no. 15, Aug. 2024, doi: 10.3390/en17153757.

[9] C. Bi, S. Xu, N. Hu, S. Zhang, Z. Zhu, and H. Yu, "Identification Method of Corn Leaf Disease Based on Improved Mobilenetv3 Model," *Agronomy*, vol. 13, no. 2, Feb. 2023, doi: 10.3390/agronomy13020300.

[10] H. Amin, A. Darwish, A. E. Hassanien, and M. Soliman, "End-to-End Deep Learning Model for Corn Leaf Disease Classification," *IEEE Access*, vol. 10, pp. 31103–31115, 2022, doi: 10.1109/ACCESS.2022.3159678.

[11] S. Z. M. Zaki, M. A. Zulkifley, M. Mohd Stofa, N. A. M. Kamari, and N. A. Mohamed, "Classification of tomato leaf diseases using mobilenet v2," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 290–296, Jun. 2020, doi: 10.11591/ijai.v9.i2.pp290-296.

[12] S. Mousavi and G. Farahani, "A Novel Enhanced VGG16 Model to Tackle Grapevine Leaves Diseases with Automatic Method," *IEEE Access*, vol. 10, pp. 111564–111578, 2022, doi: 10.1109/ACCESS.2022.3215639.

[13] Q. Dai *et al.*, "Citrus Disease Image Generation and Classification Based on Improved FastGAN and EfficientNet-B5," *Agronomy*, vol. 13, no. 4, Apr. 2023, doi: 10.3390/agronomy13040988.

[14] S. Saleem, M. I. Sharif, M. I. Sharif, M. Z. Sajid, and F. Marinello, "Comparison of Deep Learning Models for Multi-Crop Leaf Disease Detection with Enhanced Vegetative Feature Isolation and Definition of a New Hybrid Architecture," *Agronomy*, vol. 14, no. 10, Oct. 2024, doi: 10.3390/agronomy14102230.

[15] S. He, P. Peng, Y. Chen, and X. Wang, "Multi-Crop Classification Using Feature Selection-Coupled Machine Learning Classifiers Based on Spectral, Textural and Environmental Features," *Remote Sens (Basel)*, vol. 14, no. 13, Jul. 2022, doi: 10.3390/rs14133153.

[16] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technologies (Basel)*, vol. 13, no. 1, p. 32, Jan. 2025, doi: 10.3390/technologies13010032.

[17] D. Zhu, J. Tan, C. Wu, K. L. Yung, and A. W. H. Ip, "Crop Disease Identification by Fusing Multiscale Convolution and Vision Transformer," *Sensors*, vol. 23, no. 13, Jul. 2023, doi: 10.3390/s23136015.

[18] P. Christakakis, N. Giakoumoglou, D. Kapetas, D. Tzovaras, and E.-M. Pechlivani, "Vision Transformers in Optimization of AI-Based Early Detection of Botrytis cinerea," *AI*, vol. 5, no. 3, pp. 1301–1323, Aug. 2024, doi: 10.3390/ai5030063.

[19] H. Wang, L. Zhang, and J. Zhao, "Application of a Fusion Attention Mechanism-Based Model Combining Bidirectional Gated Recurrent Units and Recurrent Neural Networks in Soil Nutrient Content Estimation," *Agronomy*, vol. 13, no. 11, Nov. 2023, doi: 10.3390/agronomy13112724.

[20] Y. Wang *et al.*, "Classification of Plant Leaf Disease Recognition Based on Self-Supervised Learning," *Agronomy*, vol. 14, no. 3, Mar. 2024, doi: 10.3390/agronomy14030500.

[21] L. Bi, G. Hu, M. M. Raza, Y. Kandel, L. Leandro, and D. Mueller, "A gated recurrent units (Gru)-based model for early detection of soybean sudden death syndrome through time-series satellite imagery," *Remote Sens (Basel)*, vol. 12, no. 21, pp. 1–20, Nov. 2020, doi: 10.3390/rs12213621.

[22] X. Ni, F. Wang, H. Huang, L. Wang, C. Wen, and D. Chen, "A CNN- and Self-Attention-Based Maize Growth Stage Recognition Method and Platform from UAV Orthophoto Images," *Remote Sens (Basel)*, vol. 16, no. 14, Jul. 2024, doi: 10.3390/rs16142672.

[23] X. Wang, S. Fang, Y. Yang, J. Du, and H. Wu, "A New Method for Crop Type Mapping at the Regional Scale Using Multi-Source and Multi-Temporal Sentinel Imagery," *Remote Sens (Basel)*, vol. 15, no. 9, May 2023, doi: 10.3390/rs15092466.

[24] O. Jouini, M. O.-E. Aoueileyine, K. Sethom, and A. Yazidi, "Wheat Leaf Disease Detection: A Lightweight Approach with Shallow CNN Based Feature Refinement," *AgriEngineering*, vol. 6, no. 3, pp. 2001–2022, Jul. 2024, doi: 10.3390/agriengineering6030117.

[25] S. M. Hassan, A. K. Maji, M. Jasiński, Z. Leonowicz, and E. Jasińska, "Identification of plant-leaf diseases using cnn and transfer-learning approach," *Electronics (Switzerland)*, vol. 10, no. 12, Jun. 2021, doi: 10.3390/electronics10121388.

[26] R. A. Ansari and T. J. Mulrooney, "Self-Attention Multiresolution Analysis-Based Informal Settlement Identification Using Remote Sensing Data," *Remote Sens (Basel)*, vol. 16, no. 17, p. 3334, Sep. 2024, doi: 10.3390/rs16173334.

[27] T. Wang, H. Xia, J. Xie, J. Li, and J. Liu, "A Multi-Scale Feature Focus and Dynamic Sampling-Based Model for Hemerocallis fulva Leaf Disease Detection," *Agriculture*, vol. 15, no. 3, p. 262, Jan. 2025, doi: 10.3390/agriculture15030262.

[28] S. Parez, N. Dilshad, N. S. Alghamdi, T. M. Alanazi, and J. W. Lee, "Visual Intelligence in Precision Agriculture: Exploring Plant Disease Detection via Efficient Vision Transformers," *Sensors*, vol. 23, no. 15, Aug. 2023, doi: 10.3390/s23156949.

[29] S. A. Shah, I. Taj, S. M. Usman, S. N. Hassan Shah, A. S. Imran, and S. Khalid, "A hybrid approach of vision transformers and CNNs for detection of ulcerative colitis," *Sci Rep*, vol. 14, no. 1, p. 24771, Dec. 2024, doi: 10.1038/s41598-024-75901-4.

[30] U Barman *et al.*, "ViT-SmartAgri: Vision Transforer and Smartphone-Based Plant Disease Detection for Smart Agriculture," *Agronomy*, vol. 14, no. 2, Feb. 2024, doi: 10.3390/agronomy14020327.

[31] Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in Proceedings of the 38th International Conference on Machine Learning (ICML), Online, Jul. 18–24, 2021, vol. 139, pp. 10347–10357. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html