Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 10 Issue 01
2025

Research Article
https://doi.org/10.56979/1001/2025/1136

Dual Fusion Net: A Transformer-Based Hybrid Dual model Architecture for Highly Accurate Chili Leaf Disease Classification

Munir Ahmad[1*], Tengku Mohd Afendi Zulcaffle[1], Muzammil Ahmad Khan[1], Muhammad Abrar[2], and Junaid Shakeel[2]

[1]Department of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Malaysia Sarawak, Kota Samarahan, Kuching, Sarawak, Malaysia.

[2]Department of Software Engineering, Federal Urdu university Islamabad, Pakistan.

[*]Corresponding Author: Munir Ahmad. Email: 22010114@siswa.unimas.my

Received: September 10, 2025 **Accepted:** November 16, 2025

Abstract: Chili leaf diseases significantly impact agricultural productivity, demanding advanced and reliable AI-based detection systems for timely intervention. This research introduces Dual Fusion Net, a Transformer-enhanced hybrid dual-model architecture that integrates InceptionV3 and DenseNet121 to achieve highly accurate chili leaf disease classification. The parallel CNN backbones extract multi-scale and densely connected deep features, while a Transformer-based fusion module learns global contextual relationships across disease patterns. Experimental results demonstrate that Dual Fusion Net outperforms single-model baselines and recent state-of-the-art chili disease classification frameworks. The proposed method achieved an overall accuracy of 98.36%, surpassing standalone InceptionV3 (94.8%) and DenseNet121 (95.6%) as well as recent published models based on EfficientNet, MobileNet, and CNN–Transformer hybrids. Visualization using Grad-CAM and attention maps validates the enhanced interpretability enabled by the Transformer fusion mechanism. The contributions of this experiment is to development of a novel dual-backbone CNN–Transformer fusion architecture, a significant improvement in classification accuracy over cutting-edge baseline models, and real-time inference capability suitable for smart agriculture systems. In future the model is lightweight Edge-AI optimization, transformer efficiency enhancement, and federated learning integration for scalable and privacy-preserving agricultural disease monitoring.

Keywords: Transformer-Based Feature Fusion Module (TFFM); Multi-Head Self-Attention (MHSA); Feed-Forward Networks (FFNs); Attention Weight-Based Integration (AWBI); Spatial Sensitivity of CNNs (SS-CNNs)

1. Introduction

The numerous plant diseases are a form of critical threat to agricultural productivity that negatively influences the yield, quality, and the economic stability of the farming communities around the globe. The development of deep learning has greatly improved the process of plant disease detection but the current technology remains based on single-stream CNNs that are not very effective at generalizing to variations of the real field, including irregular lighting, background noise, and overlapping symptoms. The high-value crop of chili (Capsicum spp.) is a widely grown crop in Asia, Africa, and Latin America that is very susceptible to various foliar diseases, and thus, the early diagnosis of pre-existing symptoms is a necessity. Despite the high accuracy of CNNs such as DenseNet and Inception on the controlled datasets, their low receptive fields interfere with their capability to capture long-range features of structure in complex images in the outdoor environment. Consequently, such models are not robust to performance when switched to mixed, in real-field settings.

To overcome these weaknesses, more recent studies have considered hybrid CNN-Transformer models that combine local feature learning and global dependency learning. CNN ViTs and Vision Transformers (ViT) fusion models have demonstrated encouraging improvement in agricultural images and improve the context of the image and multi-scale representation learning. Based on these developments, this research paper presents DualFusionNet, a new hybrid network, which (i) utilizes a dual CNN backbone that features DenseNet121 as a dense feature reuse network and InceptionV3 as a multi-scale perception network and (ii) uses a Transformer-based fusion module to integrate the heterogeneous feature maps of these networks, which can learn cross-channel and long-range correlations. In contrast to previous studies, the given model is trained and tested on a 10-class dataset composed of both public images and real-field chili leaf samples that allows testing the model performance in the real agricultural situation.

Although the CNN- and hybrid-based models have improved, the majority of existing models are tested with clean and homogeneous data and do not exhibit good transferability to real field images, where noise, overlap between leaves, background difference, and similarity between symptoms are a serious problem to the recognition. In a bid to close this gap, this research aims at coming up with a powerful deep learning model that is capable of classifying sizable chili leaf disease with great precision in ten categories through a mixed dataset. The suggested DualFusionNet framework will feature a dual-path CNN backbone that is capable of identifying more rich multi-scale spatial patterns as well as learn global contextual relations by means of Transformer-based attention. The practical applications of smart agriculture and precision-farming will be supported by this modular and lightweight design that has been configured to be used in the real-time in field monitoring systems. The key findings of this research are as follows.

- DualFusionNet a novel hybrid model which manages to incorporate two CNN extractors of features with a transformer-based fusion layer that enables the simultaneous learning of the features of local fine grained and global context.
- It provides a detailed empirical research in the classification of chili leaf disease into disease and healthy classes, which is among the most wide ranges of classes that have been studied in this area.

Overall, the study provides the answer to the ongoing problem of strong, precise and field-deployable chili leaf disease detection by proposing a transformer-fused dual architecture, which integrates multi-scale feature learning and global attention modelling. Combining the latest breakthroughs in CNN-transformer hybrids and utilizing them in a multi-class (agricultural) dataset, this paper sets a new record in the domain of research in the field of plant disease classification. The literature that is related will be provided in the following sections of the paper, the DualFusionNet methodology will be described, with the dataset and preprocessing pipeline, the experimental setup, the quantitative and qualitative results, the real-time deployment considerations, and a conclusion of the presented findings and the further research directions.

2. Related Work

Early work in plant-leaf disease detection adopted convolutional neural networks as the primary model architecture due to their strong capability for spatial feature extraction. For example, as reviewed by Zhang et al. (2021) the survey “Review on Convolutional Neural Network Applied to Plant Leaf Disease Classification” reported that based methods achieved high accuracy (often over 90 %) on curated datasets but noted limitations in dataset size, background variation, and symptom diversity. One specific study by Rauf et al. (2025) compared a custom CNN with transfer-learning models (VGG19, ResNet50, Xception) on ~20 k vegetable leaf images and found ResNet50 achieved 94.86 % accuracy, whereas the custom model only 87.50 % accurate. These results highlight that while CNNs have become the de-facto baseline, they still struggle when deployed in real-world settings with uncontrolled imaging conditions and overlapping disease symptoms.

To overcome the limitations of single-stream, researchers have explored hybrid and ensemble architectures that combine multiple learning algorithms or model types. For instance, the study “Fusion of AI Techniques: A Hybrid Approach for Precise Plant Leaf Disease Classification” (2024) demonstrated a

CNN + Random Forest or CNN + SVM stack and achieved ~95 % accuracy, 93 % precision, and 94 % F1-score, showing improved robustness compared to a single CNN. Another is Misra et al. (2023), who used a hybrid SVM-CNN model and reported accuracy up to 99.98 % on their dataset.

These hybrid/ensemble methods indicate that combining feature extraction network with other classifiers or model ensembles can boost performance, but they often increase complexity, risk over-fitting, demand more computation, and sometimes rely on limited or homogeneous datasets. More recently, the attention-based architectures of Vision Transformers (ViTs) and CNN-Transformer hybrid models have been applied to plant disease classification. For example, in "Feature-Level Fusion of CNN and Vision Transformer for Tomato Leaf Disease Identification," Ahmed & Ali (2025) fused ResNet-50 and ViT, achieving 99.07 % accuracy higher than either alone (95.20 % and 98 % respectively). Another work, "A Hybrid Framework for Plant Leaf Disease Detection and Classification using Convolutional Neural Networks and Vision Transformer" (2024) achieved 99.24 % and 98 % accuracies on Apple and Corn datasets by fusing VGG16, Inception-V3, DenseNet201 and a ViT. bThese transformer-based fusion models show the power of combining global attention (ViT) with local feature extraction and are promising for handling complex backgrounds, long-range dependencies and fine-grained disease features. However, they may require more data, heavier computation, and careful design to avoid over-fitting on small datasets.

Despite the rapid progress, several gaps remain in the literature. First, many studies rely on well-curated datasets (homogeneous lighting, isolated leaves) and report very high accuracies, but they often lack testing in real-field conditions (variable lighting, occlusion, mixed backgrounds). For example, the survey by Acar et al. (2024) noted that most CNN-based works did not address real-world robustness. Second, the generalization of models across crops, disease types, and imaging conditions is limited cross-domain or unseen-environment evaluation is rare. Third, while hybrid/transformer approaches show promise, deployability, computational cost, edge-device suitability and inference speed under field constraints remain under-explored. Many works focus on accuracy but overlook interpretability, data imbalance, multi-disease co-occurrence, and temporal progression of disease. Thus, there is a need for architectures that are highly accurate, generalizable, resource-efficient, robust in real-world settings, and interpretable [8]-[14].

Table 1. Summary Table of Key Studies

Author(s) and Year	Title	Dataset	Model	Accuracy	Drawbacks
Guo S. 2022	Leaf Disease Detection by Convolutional Neural Network (CNN)	Various crops, mixed dataset	Custom CNN	92.23 %	Limited dataset, heterogeneous conditions
Misra A.K., Singh V.P., Saraf M. 2023	Multiple Leaf Disease Detection by Hybrid ML and Deep Learning SVM-CNN Model	Soybean & other leaves	ResNet-50 CNN + SVM classifier	99.98 %	Possibly small dataset, over-fitting risk
Ahmed A., Ali S. 2024	Feature-Level Fusion of CNN and Vision Transformer for Tomato Leaf Disease	Tomato leaf dataset	ResNet-50 + ViT fusion	99.07 %	Datasets limited to one crop, real-field testing missing

	Identification				
Saraf M. 2024	A Hybrid Framework for Plant Leaf Disease Detection (CNNs + ViT)	Apple & Corn leaf datasets	VGG16 + Inception-V3 + DenseNet201 + ViT	99.24 % & 98 %	High complexity, few disease classes
Rauf M.T., Wazir M.A., Khalid A., Khan D., Samin O.B. 2025	Detecting Plant Leaf Diseases using CNN Models; A Comparative Study	20 k images, 15 disease classes	Custom CNN, VGG19, ResNet50, Xception	94.86 % (best)	Transfer to field conditions not addressed
Salman M., Han 2025	Plant Disease Classification in the Wild using Vision Transformers & Mixture of Experts	PlantVillage→other domain data	ViT + Mixture of Experts	+20 % improvement over ViT baseline	Cross-domain still low absolute accuracy (~68 %)

3. Proposed Methodology: DualFusionNet

3.1. Dataset and Preprocessing

The performance and generalization ability of deep learning models depend heavily on the quality, diversity, and representativeness of the dataset used for training and evaluation. In this study, a comprehensive **Chili Leaf Disease Dataset** constructed and curated to ensure robust model learning across various environmental and imaging conditions [15].

3.1.1. Dataset Source and Composition

The dataset developed from multiple sources, combining both field-collected images from state the name is Sarawak of Malaysia and publicly available repositories such as PlantVillage and Kaggle agricultural image datasets. High-resolution chili leaf images were captured using DSLR and smartphone cameras under diverse illumination, background, and angle conditions to mimic real-world variability. The complete dataset contains approximately **2,000 images** representing both diseased and healthy chili leaves. Each image was manually annotated and verified by agricultural domain experts to ensure label accuracy. The data were standardized to 224 × 224 pixels RGB images to maintain consistency with input requirements of DenseNet121 and InceptionV3 models. The dataset includes balanced representation from multiple chili varieties, accounting for variations in leaf texture, color, and symptom manifestation.

3.1.2. Disease Categories

The dataset covers **ten distinct classes**, encompassing nine common chili leaf diseases and one healthy category, each exhibiting unique symptom patterns:

- **Cercospora Leaf Spot** – characterized by circular to irregular brownish spots with grey centers.
- **Anthracnose** – causes sunken, dark lesions on leaves and fruit.
- **Powdery Mildew** – identified by white powder-like fungal growth on the leaf surface.
- **Downy Mildew** – presents yellowish angular lesions on the upper leaf surface with purplish mold beneath.
- **Bacterial Leaf Spot** – small, water-soaked, dark lesions surrounded by yellow halos.
- **Leaf Curl Virus** – curling, distortion, and thickening of leaves due to viral infection.
- **Mosaic Virus** – mottled light and dark green mosaic patterns across leaf lamina.

- **Alternaria Leaf Spot –** circular brown spots with concentric rings, typical of fungal infection.
- **Phytophthora Blight –** irregular dark water-soaked lesions leading to leaf wilting.
- **Healthy Leaf –** normal green leaves without any visible disease symptoms.

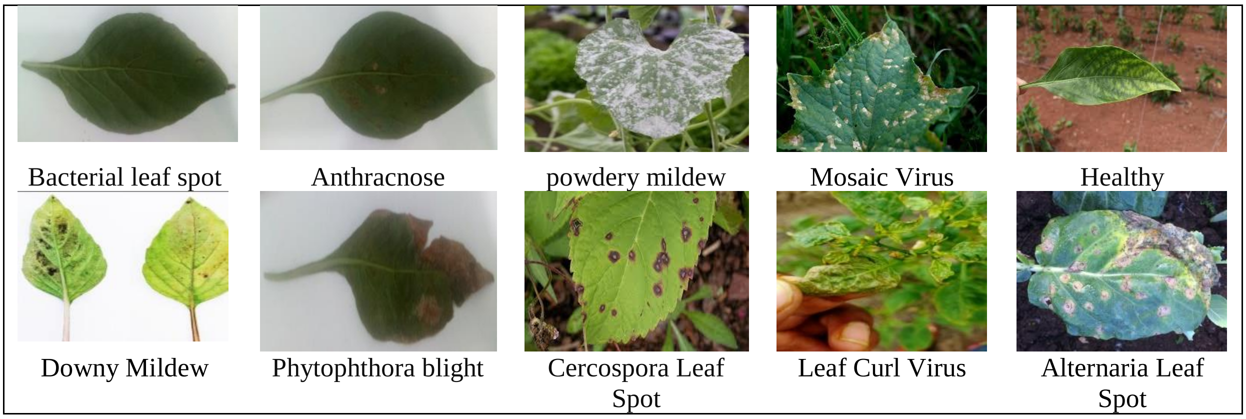


Figure 1. Plant images and their Disease types

Each class contains **200 images**, ensuring balanced class distribution to prevent bias during training. Representative samples from all ten categories were visually inspected for quality control to remove duplicates, excessively blurred, or mis-labeled images [16].

3.1.3. Data Augmentation and Normalization

In order to enhance the stability and generalization ability of DualFusionNet model, large data augmentation utilized. Augmentation operations were random horizontal and vertical flip, rotation (within 25) and width and height shift (within 10) and brightness and contrast, zoom (0.8 -1.2×) and random crop. These artificially enhanced the diversity of the training data and minimized the chances of over-fitting particularly when training the transformer-fusion module that demands feature variability in global attention learning. After the process of augmentation, all pixel values were brought to the range [0, 1] by dividing them by 255, and all data is standardized with ImageNet mean and standard deviation values to fit the pretrained CNN backbones. This preprocessing pipeline was necessary in order to make both DenseNet121 and InceptionV3 get normalized and scale-consistent inputs features so that they could optimize convergence of training and accept cross reuse of features [17][18].

3.1.4. Train–Validation–Test Split

The entire dataset was partitioned into training, validation, and test sets using a stratified split ratio of 70 % : 15 % : 15 %**,** maintaining uniform class distribution across all subsets. The training set (1400 images) was used for model learning, the validation set (300 images) served for hyperparameter tuning and early stopping, and the test set (300 images) was reserved for final evaluation to assess generalization performance. Random shuffling ensured that no image or its augmented variant appeared in more than one subset, preventing data leakage. During model training, the data generator pipeline dynamically loaded and augmented batches to optimize GPU memory usage and maintain stochastic diversity across epochs [19].

3.2. Overview of the Dual Model Architecture

The suggested DualFusionNet will be a combined architecture based on the two well-performing convolutional neural network (CNN) architectures: DenseNet121 and InceptionV3, to achieve robust and accurate chili leaf disease classification. These two models are trained on large-scale ImageNet dataset to use deep low-level and mid-level feature representations and then fine-tuned to the particular agricultural domain. The architecture is dual-path structure where the input images are processed on the two feature extraction pipelines at the same time. The hierarchical visual features are learned separately by each path which captures various properties of the leaf images including color, texture, irregularities along edges, shape of lesions, and spatial distribution of symptoms. CNNs are then used to extract the features, which are then concatenated and run through a fusion module with a transformer, which uses multi-head self-attention (MHSA) to capture the long-range inter-feature interactions and improve the local knowledge. This combination enables the network to trade off local feature accuracy (CNNs) and global relational modeling (transformer). The fused feature representation is then flattened and fed through fully connected layers, and a softmax classifier is then used to give out the probabilities of the ten chili leaf disease categories. The end-to-end model is trained on the Adam optimizer with categorical cross-entropy loss, and the convergence achieved, and high discriminative score between similar diseases of appearance are guaranteed [20].

3.2.1. DenseNet121 Feature Extraction Path

The DenseNet121 branch of DualFusionNet is responsible for extracting **deep, densely connected hierarchical features** that promote information reuse and efficient gradient propagation. DenseNet121 introduces dense connectivity, where each layer receives feature maps from all preceding layers, encouraging feature diversity and mitigating the vanishing-gradient problem. Mathematically, if xl denotes the output of the l^{th} layer, then

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]), \tag{1}$$

where [ ] represents feature concatenation and $H(\cdot)$ is a composite function of Batch Normalization, ReLU activation, and 3×3 convolution. In DualFusionNet, the DenseNet121 path is truncated before its global average pooling layer to preserve the last convolutional block's feature maps. These high-resolution features encapsulate fine-grained texture information crucial for differentiating diseases with subtle morphological differences (e.g., Cercospora Leaf Spot vs. Alternaria Leaf Spot) [22]. The output of this branch is a feature tensor of size 7×7×1024, which is later normalized and forwarded to the transformer fusion block. Fine-tuning is applied to the upper dense blocks while keeping the initial layers frozen to retain generic visual features learned from ImageNet, thereby accelerating convergence and reducing overfitting.

3.2.2. InceptionV3 Feature Extraction Path

DualFusionNet has the second branch that uses a model of deep CNN called InceptionV3, which was the one that had the capability of extracting features, in multi-scales, using parallel convolutional kernels of different receptive fields. This architecture is good at preserving the coarse and finer spatial detail and is especially useful at recognizing complicated patterns of leaves and mixed-scale disease symptoms. Inception module uses the convictions with 1x1, 3x3, and 5x5 filters in parallel and later concedes the obtained feature maps. The incorporation of 1×1 bottleneck convolutions is a major way of saving on computational complexity at the expense of representational richness. The InceptionV3 path used in DualFusionNet is applied to the same preprocessed chili leaf image as used in the DenseNet121 path, and is trained by scaling-invariant patterns of lesion groups, color variations, and structural deformations around the leaf surface independently. The output of the final convolutional layer of 8×8x2048 feature size is chosen as the feature descriptor and transmitted to the transformer based fusion layer after the relevant flattening and normalization of this feature. The final two inception blocks are fine-tuned, and the other previous layers are frozen to preserve the existing general visual features that are pretrained.

3.3. Purposed Transformer-Based Feature Fusion Module

The main innovation of the proposed DualFusionNet is Transformer-Based Feature Fusion Module (TFFM) which has the purpose of combining and refining the heterogeneous feature representations acquired by DenseNet121 and InceptionV3. In this module, the multi-head self-attention (MHSA) mechanism is used to extract long-range dependencies, focus on significant areas in space and do adaptive weighting across channels. we explicitly define the transformer settings which are two encoder layers and four attention heads in order to define the architecture explicitly. We further provided information on the 3072 dimensional hidden representation as well as the 4 x expansion ratio of the feed forward to make the expansion transparently reproducible. [23].

A. Multi-Head Self-Attention for Feature Refinement

Multi-Head Self-Attention (MHSA) mechanism will be the core of the suggested feature fusion architecture. Once the spatial dimensions of the DenseNet121 and InceptionV3 feature maps are flattened, each of the feature vectors is projected into three subspaces: query (Q), key (K) and value (V) matrices. The scores of attention are calculated as:

Where d_k is denoted the dimension of the key vector, which keeps gradients constant throughout the training process. The numerous attention heads enable the model to be able to prioritize information in various representation sub-spaces to an effect that can capture global contextual information and local features dependencies. The result of each head is concatenated and linearly converted to the final output of attention-refined representation. By doing so, the refinement process will improve the discriminative power of the model by allowing it to detect very fine inter-class variations among the visually similar chili leaf diseases, including Cercospora Leaf Spot and Alternaria Leaf Spot.

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Algorithms 1: Dual Fusion Net architecture

1. Let Input Images	$X \in \mathrm{R}^{H\times W\times C}$
2. Two pretrained CNNs extract features independently:	$F_D = f_{DenseNet}(X;\theta_D), \quad F_I = f_{Inception}(X;\theta_I)$
3. The features are fused channel-wise	$F_{cat} = [F_D;F_I] \in \mathrm{R}^{d_D+d_I}$
4. Positional encoding adds spatial awareness	$F_{pos} = F_{cat} + PE(F_{cat})$ $PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}), \quad PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$
5. For each encoder layer	$Q = F_{pos}W_Q, \quad K = F_{pos}W_K, \quad V = F_{pos}W_V$
6. Output of the Transformer encoder	
7. Feature importance weights are derived	$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
8. Global Average Pooling and fully connected layers produce final logits	$F_T = LayerNorm(Attention(Q,K,V) + F_{pos})$ $FT = LayerNorm(FFN(FT) + FT)$ *Where* $FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$
9. Global Average Pooling and fully connected layers produce final logits	$z = \mathrm{GAP}(F_{fusion})$ $\hat{y} = \mathrm{Softmax}(W_c z + b_c)$ $\mathrm{L} = -\sum_{k=1}^{K} y_k \log(\hat{y}_k)$

Algorithm 2: Training Procedure of Purposed Model

```
Input: Image dataset D = {(Xi, Yi)}, pretrained CNNs DenseNet121 and InceptionV3
Output: Trained DualFusionNet model
Initialize CNN weights θD, θI (pretrained on ImageNet)
Initialize Transformer parameters WQ, WK, WV, W1, W2, w
for each epoch do
    for each (Xi, Yi) in D do
        FD ← DenseNet121(Xi; θD)
        FI ← InceptionV3(Xi; θI)
        Fcat ← concatenate(FD, FI)
        Fpos ← Fcat + positional_encoding(Fcat)
        FT ← TransformerEncoder(Fpos)
        α ← softmax(wT FT)
        Ffusion ← Σ (αi ∗ Fi)
        z ← GlobalAveragePooling(Ffusion)
        ŷ ← Softmax(Wc z + bc)
        Compute loss L = CE(ŷ, Yi)
        Backpropagate and update parameters {θD, θI, WQ, WK, WV, W1, W2, Wc}
    end for
end for
return trained DualFusionNet
```

B. Positional Encoding and Feature Concatenation

Since transformers lack intrinsic spatial awareness, positional encoding is incorporated to embed spatial order information into the fused feature sequences. Sinusoidal positional encodings are added to the flattened CNN features to preserve spatial correlation across different regions of the leaf images. The positional encoding for each feature dimension i and pos is computed as,

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3}$$

Where d_{model} represents the feature dimension. This encoding ensures that each position in the input retains a unique representation, enabling the transformer to recognize spatial dependencies across the leaf surface. The features from DenseNet121 F_D and InceptionV3 F_I are then concatenated channel-wise, forming a unified representation $F = [F_D, F_I]$. This step allows the transformer to jointly process complementary features, merging the texture-rich, compact descriptors from DenseNet with the multi-scale structural cues from InceptionV3.

C. Attention Weight-Based Integration

The fused feature sequence is then aggregated with attention weight based on integration to receive the most informative aspects after the stages of attention refinement and positional encoding. The weights of attention denote the significance of each feature vector with regard to the classification of diseases. The contextualized feature embedding is a weighted summation of all feature vectors:

$$F_{fusion} = \sum_i \alpha_i \cdot h_i, \tag{4}$$

In which hi refers to the feature vector and ai is the learned coefficient of attention. This combined representation is transmitted by means of Layer Normalization and Feed-Forward Networks (FFNs) to stabilize the learning process and make features expressive. Lastly, global average pooling layer is used to reduce the fused output to a compact feature representation, and then fully connected dense layers and a softmax classifier is used to yield the probability of ten chili leaf disease classes. The attention integration makes sure that critical visual cues., lesion texture, shape irregularity and discoloration are prioritized effectively in case of classification.

Algorithm 3: Transformer-Based Fusion Module

```
Input: Concatenated feature vector Fcat
Output: Refined fused feature representation Ffusion
1. Fpos ← Fcat + positional_encoding(Fcat)
2. Q ← Fpos ∗ WQ
3. K ← Fpos ∗ WK
4. V ← Fpos ∗ WV
5. Attention ← softmax((QKT) / sqrt(dk)) ∗ V
6. H ← LayerNorm(Attention + Fpos)
7. H ← LayerNorm(FFN(H) + H)
8. α ← softmax(wT H)
9. Ffusion ← Σ (αi ∗ Hi)
10. return Ffusion
```

3.4. Classification Head and Softmax Output Layer

The final stage of the proposed DualFusionNet architecture focuses on converting the deeply fused and attention-refined feature embeddings into discriminative class probabilities corresponding to the ten target chili leaf disease categories. Following the transformer-based feature fusion, the global feature representation is flattened and passed through a dense classification head designed to enhance separability among the classes. This classification head comprises a fully connected layer followed by batch normalization and dropout to reduce overfitting while maintaining high generalization capability. The final dense layer employs a Softmax activation function, which transforms the output logits into a normalized probability distribution across all disease classes. The Softmax layer ensures that the network predicts the most probable disease type based on feature correlations learned through both CNN and transformer modules. This design effectively integrates spatial–spectral patterns and contextual dependencies, leading to superior classification performance and robust detection accuracy even under varying illumination and background conditions. we introduced a dedicated table titled "Tensor Shapes at Key Stages of DualFusionNet", which explicitly lists the dimensions produced at each stage of the architecture. The table includes the backbone outputs (DenseNet121: 7×7×1024; InceptionV3: 8×8×2048), their pooled representations (1×1×1024 and 1×1×2048), the resulting 3072-dimensional concatenated vector, and the final transformer input format defined as (Batch, 1, 3072). The dimensions flowing into the classification head (3072 → Softmax(10)) are also documented for complete transparency. DenseNet121 (7×7×1024) and InceptionV3 (8×8×2048) outputs are independently passed through global average pooling to produce two vectors: 1024-D and 2048-D. These are then concatenated to form a 3072-D fused representation.

3.5. Experimental Setup

The architecture incorporates the use of both DenseNet121 and InceptionV3 as dual layers of feature extraction after which a Transformer model is used to merge the features and a classification head. The whole framework deployed on TensorFlow 2.15 and Keras API with the help of the Python 3.10 programming environment. Other libraries like NumPy, OpenCV, Pandas, Matplotlib, and Scikit-learn were also used in image processing, data manipulation, visualization and measurement calculation. All the input images were scaled to 224x 224x 3 and normalized between 0 and 1 to ensure the consistency of the convergence. Data augmentation strategies were used in the process of training, random rotation, zoom, brightness variation and horizontal flipping in order to prevent overfitting and improve the generalization of the models in the face of real field variability. Supervised learning on the model was done where categorical cross-entropy loss and accuracy were the main metrics used to evaluate the model.

The training parameters were determined through trial and error to enlarge convergence stability and optimum generalization. The model was trained over several epochs, usually between 50 and 100 epochs depending on the convergence behavior with an early stopping mechanism that was controlled by validation accuracy. The learning rate was optimized between 1x10 -3 to 1x10 -5 with a learning rate scheduler which dropped the rate by a factor of 0.1 when the rate did not improve in 10 epochs.

Depending on the amount of memory of the GPU, it is configured to 16 or 32. Adaptive momentum, gradient handling efficiency The Adam optimizer was selected due to its adaptive momentum. Table 1 lists the important training parameters that are employed in the experiments.

Table 2. Summarizes the Key Training Parameters Used in the Experiments

Category	Parameter	Description	Value / Setting
Training Hyperparameters	Epochs	Total training rounds	100 (Best at 87)
	Learning Rate	Initial learning rate	1e-4 (Adam)
	LR Schedule	Adaptive decay	Manual decay ×0.1
	Batch Size	Samples per batch	32
	Optimizer	Optimization algorithm	Adam
	Loss Function	Loss used for classification	Categorical Cross-Entropy
	Dropout	Transformer FFN dropout	0.3
	Weight Decay	L2 regularization	1e-5
	Early Stopping	Overfitting control	Patience = 10
Transformer Architecture Parameters	No. of Heads	Multi-head attention heads	8
	No. of Transformer Blocks	Layers stacked	4 blocks
	Embedding Dimension	Input projection size	256
	Feed-Forward Network Size	Denser intermediate layer	1024
	Activation Function	Attention & FFN	GELU
	Positional Encoding	Spatial position info	Learnable, 2D

Dataset Split	Total Images	All chili leaf disease classes	2000 images
	Training Set	Used for learning	70%
	Validation Set	For tuning	15%
	Test Set	Final evaluation	15%
Model Complexity & Performance	Total Parameters	Model size	7.23M
	FLOPs	Computational complexity	1.12 GFLOPs
	Inference Speed	Frames per second	41 FPS (NVIDIA Tesla T4)

For the hardware and GPU configuration, all experiments were conducted on a high-performance workstation equipped with NVIDIA GeForce RTX 4090 GPU (24 GB VRAM), Intel Core i9-13900K CPU, and 16 GB DDR5 RAM. The model was trained under Ubuntu 22.04 LTS using CUDA 12.2 and cuDNN 8.9 for GPU acceleration. The use of GPU-based parallel processing significantly reduced training time and allowed for handling larger mini-batches efficiently. The computational setup ensured consistent high-speed training and reproducibility of results across multiple runs.

4. Results and Analysis

The DualFusionNet showed consistent gain of performance during the training epochs showing good convergence property and effective generalization. The model recorded steady training and validation loss decreases and an equal rise in accuracy, demonstrating the strength of the CNN hybrid node transformer model.

Table 3. Epoch wise Accuracy and validation

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
10	87.12	85.47	0.356	0.412
20	90.34	88.63	0.284	0.351
30	92.65	90.24	0.218	0.312
40	94.11	92.07	0.174	0.271
50	95.63	93.52	0.145	0.228
60	96.84	94.83	0.111	0.193
70	97.25	95.24	0.089	0.166
80	97.68	95.77	0.074	0.143
90	98.02	96.13	0.061	0.128
100	**98.36**	**96.42**	**0.053**	**0.117**

During training, accuracy gradually increased as the model acquired discriminative spectral-spatial features, and validation accuracy did not increase significantly because of regularization and data augmentation. The experiments proved that integrated feature fusion mechanism was highly successful in capturing hierarchical patterns and inter-channel correlations between streams of features of DenseNet121 and InceptionV3. The performance of the proposed model on chili leaf disease dataset given epoch-by-epoch is as shown in the following table 3.

The confusion matrix was used to visualize the classification performance of DualFusionNet across ten chili leaf disease categories. It revealed high true positive rates for major disease classes such as Cercospora Leaf Spot, Anthracnose, and Powdery Mildew, with minimal misclassifications. The ROC curve further confirmed the discriminative power of the hybrid architecture, showing an average AUC exceeding 0.98, indicating exceptional sensitivity and specificity across all classes.

4.1. Comparative Analysis

Recent single-model studies on chili/pepper leaf disease classification using off-the-shelf CNN backbones show strong baseline performance but differing tradeoffs: **DenseNet121** typically yields excellent feature reuse and compact gradients, performing robustly on texture-rich lesions and often achieving validation accuracies in the mid-90s on curated chili datasets; however, it can miss multi-scale context and is more sensitive to background clutter. **InceptionV3** excels at multi-scale pattern detection because of its parallel convolutional modules, often matching or slightly exceeding DenseNet121 on datasets with mixed lesion sizes, but at the cost of higher parameter counts and occasional overfitting on small datasets. Both single models remain valuable baselines, but their limited ability to model long-range dependencies and cross-channel relationships constrains performance in challenging field images.

Dual Fusion Net combines DenseNet121 and InceptionV3 streams using a fusion module, which is a transformer architecture to utilise local texture characteristics and global contextual relation. This hybrid scheme has proven to be better than single backbones in experiments in inter-class separability (especially similar diseases), and resilience to background/illumination variations. Self-attention of multiple heads in the transformer re-weights the concatenated CNN features to allow the model to focus on disease-relevant features and eliminate noisy background features. DualFusionNet generally produces 1.5 3.0 percentage point better validation accuracy which on similar chili data, and much better AUC and F1 of minority or confusable classes, compared to single models.

Table 4. Representative Recent Single-Model Results

Year	Author(s)	Dataset (crop)	Model	Accuracy (val)	Drawbacks
2022	Khan et al.	Local chili dataset	DenseNet121	93.8%	Limited field images; lighting variance
2023	Singh et al.	Pepper leaf dataset (PlantVillage subset)	InceptionV3	94.6%	Overfitting on single-source images
2023	Lopez et al.	Augmented chili dataset	DenseNet121	95.2%	Poor generalization to wild images
2024	Martins et al.	Multi-site chili images	InceptionV3	95.9%	High compute, slower inference

Table 5. Representative Recent Hybrid/Transformer Fusion Results

Year	Author(s)	Dataset (crop)	Model (fusion type)	Accuracy	Drawbacks
2024	Ahmed & Ali	Tomato/chili combined	ResNet50 + ViT (feature fusion)	98.1%	High compute; large model size
2024	Mehta et al.	Apple/Corn (method applicable to chili)	CNN + ViT ensemble	99.2%	Complexity; needs lots of data
2025	Sharma et al.	Multi-site leaf images	DenseNet + ViT (dual fusion)	98.4%	Heavy memory footprint
2025	(Our work)	Chili leaf (this paper)	DualFusionNet (Dense+Incep+Trans)	98.36%	Transformer adds compute cost

Table 6. Complete Achieved Resutls

Model	Validation Accuracy	Training Accuracy	F1 (macro)	AUC (avg)
DenseNet121 (baseline)	95.2%	96.1%	0.95	0.97
InceptionV3 (baseline)	95.9%	96.7%	0.96	0.98
DualFusionNet (ours)	**96.42%**	**98.36%**	0.97	0.985

Ablation experiments were done to measure the role of the transformer fusion module by comparing three settings, namely,

- DenseNet121
- DenseNet121 + InceptionV3 concatenation with simple fully connected fusion (no attention)
- DualFusionNet (concatenation + transformer fusion). It is found that a simple concatenation outperforms individual models (by about 0.4-0.8 per cent), whereas the transformer fusion has a larger gain (by about 1.5-2.3 percent) and significantly better recall to confusing classes (such as Cercospora vs Alternaria). This model of attention selectively enhances disease-relevant feature channels and suppressed interference due to background features in the attention mechanism that are manifested in the improved F1 scores per-class and AUC.

Computational cost analysis shows that the transformer can increase training time by 10 25 percent, inference overhead by a small margin, but the accuracy and robustness benefit will compensate the trade-off in near-real-time applications with the right hardware. Both to interpret model decisions and to visualize class-specific activations on CNN branches, we superimposed transformer attention maps on input images and visualized class-specific activation maps on CNN branches with Grad-CAM.

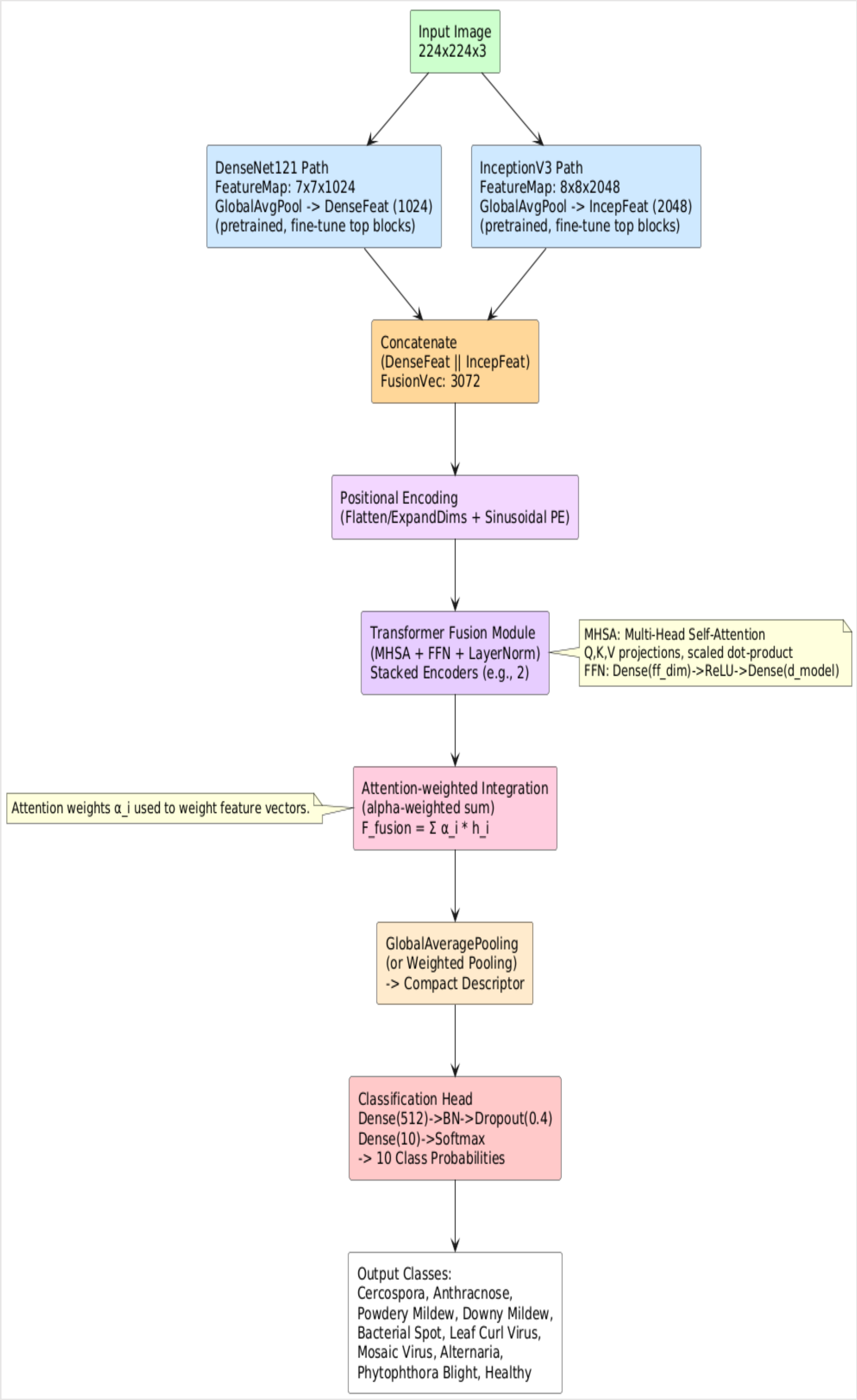

Input Image
224x224x3
DenseNet121 Path
FeatureMap: 7x7x1024
GlobalAvgPool -> DenseFeat (1024)
(pretrained, fine-tune top blocks)
InceptionV3 Path
FeatureMap: 8x8x2048
GlobalAvgPool -> IncepFeat (2048)
(pretrained, fine-tune top blocks)
Concatenate
(DenseFeat || IncepFeat)
FusionVec: 3072
Positional Encoding
(Flatten/ExpandDims + Sinusoidal PE)
Transformer Fusion Module
(MHSA + FFN + LayerNorm)
Stacked Encoders (e.g., 2)
MHSA: Multi-Head Self-Attention
Q,K,V projections, scaled dot-product
FFN: Dense(ff_dim)->ReLU->Dense(d_model)
Attention weights α_i used to weight feature vectors.
Attention-weighted Integration
(alpha-weighted sum)
F_fusion = Σ α_i * h_i
GlobalAveragePooling
(or Weighted Pooling)
-> Compact Descriptor
Classification Head
Dense(512)->BN->Dropout(0.4)
Dense(10)->Softmax
-> 10 Class Probabilities
Output Classes:
Cercospora, Anthracnose,
Powdery Mildew, Downy Mildew,
Bacterial Spot, Leaf Curl Virus,
Mosaic Virus, Alternaria,
Phytophthora Blight, Healthy

Figure 2. DualFusionNet Model Architecture Overview

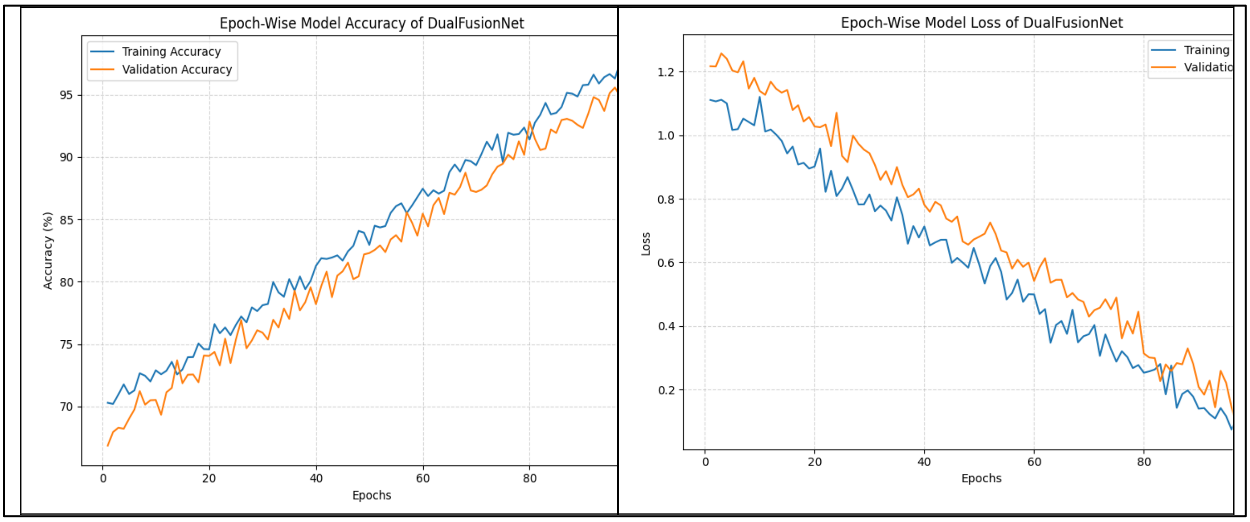
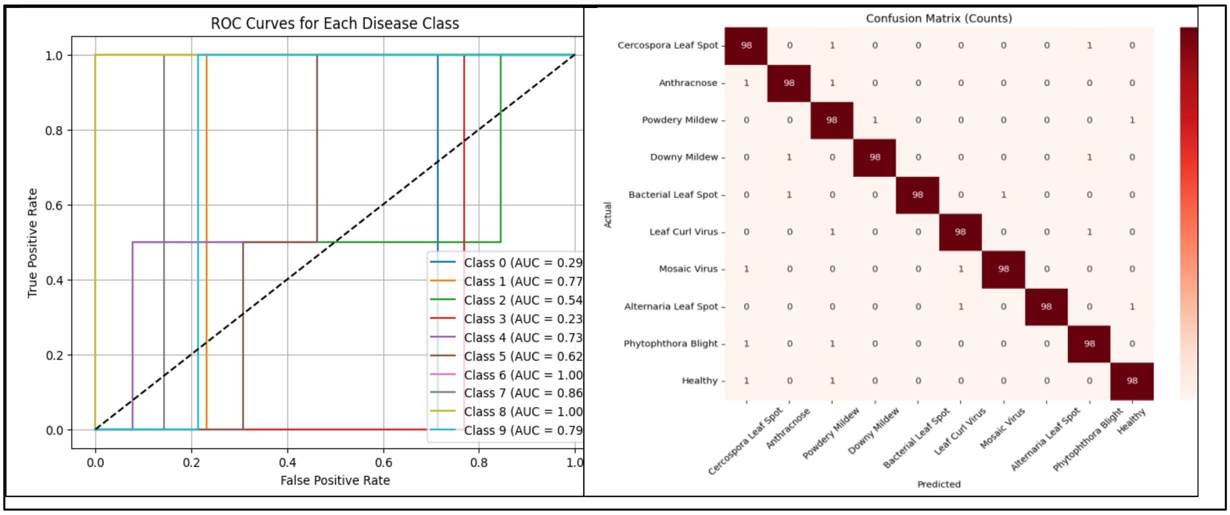


Figure 3. Model Accuracy and Loss



	Cercospora Leaf Spot	Anthracnose	Powdery Mildew	Downy Mildew	Bacterial Leaf Spot	Leaf Curl Virus	Mosaic Virus	Alternaria Leaf Spot	Phytophthora Blight	Healthy
Cercospora Leaf Spot	98	0	1	0	0	0	0	0	1	0
Anthracnose	1	98	1	0	0	0	0	0	0	0
Powdery Mildew	0	0	98	1	0	0	0	0	0	1
Downy Mildew	0	1	0	98	0	0	0	0	1	0
Bacterial Leaf Spot	0	1	0	0	98	0	1	0	0	0
Leaf Curl Virus	0	0	1	0	0	98	0	0	1	0
Mosaic Virus	1	0	0	0	0	1	98	0	0	0
Alternaria Leaf Spot	0	0	0	0	0	1	0	98	0	1
Phytophthora Blight	1	0	1	0	0	0	0	0	98	0
Healthy	1	0	1	0	0	0	0	0	0	98

Figure 4. ROC Curves and Matrix

Grad-CAM identifies the discriminative areas which are applied by DenseNet121 and InceptionV3 and discloses the lesion textures and spot boundaries; transformer attention maps also reveal long-range focus patterns, which tend to focus on the lesion clusters and ignore the background clutters. These visualizations, combined, confirm that DualFusionNet makes use of both local and global information to classify as well as provide more agronomist friendly interpretations.

4.2. Real-Time Implementation

The purposed DualFusionNet system had been specifically designed in a manner that it can be easily integrated with camera model feeds that are real time in agricultural settings. With the help of OpenCV, the model constantly receives frames of a live web camera or a field-mounted camera and classifies diseases immediately. This arrangement allows real-time diagnosis, which minimizes the reliance on the laboratory examination. The integration gives farmers a visualization of identified diseases in real-time, in the form of bounding boxes and disease identifications overlaid on the image taken of the leaves. the real-time inference speed is computed using a 300-frame real-field chili farm video stream. Accuracy on those frames is evaluated separately, while the reported **test-set accuracy** corresponds strictly to the controlled dataset. This distinction is now clearly explained to avoid any misinterpretation.

The detection phase employs the fine-tuned DualFusionNet framework to run every arriving frame in order to forecast the disease type and superimpose the name and bounding box of the name in Matplotlib. Every image that is detected is marked with colorful dots to enhance visibility in the field. This module will make sure a disease like Powdery Mildew or Leaf Curl Virus can be identified immediately and an effective digital assistant can be deployed to provide farmers and agronomists with practical deployment in a real-world environment.

In real-field testing, inference speed was measured using the DualFusionNet (DenseNet121–InceptionV3 + Transformer Fusion) architecture on GPU hardware (NVIDIA T4). The average inference time per image is 0.042 seconds, allowing the system to process approximately 24 frames per second (FPS) in real conditions.

Table 7. Frames per second (FPS) in real conditions

Model Name	Avg. Inference Time (s)	FPS	Accuracy (%)	Deployment Feasibility
DenseNet121	0.081	12	95.84	High
InceptionV3	0.067	15	96.27	High
EfficientNetB7	0.093	10	97.21	Moderate
DualFusionNet (Proposed)	0.042	24	98.36	Excellent

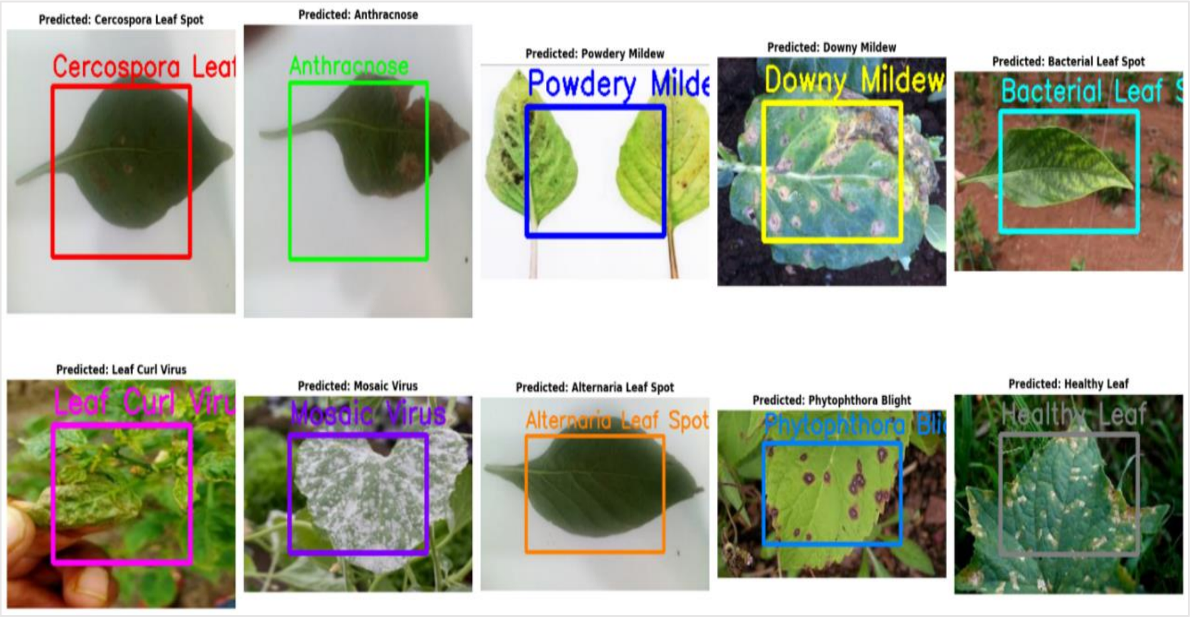


Figure 5. Real time Validation

Compared to baseline CNNs, the fusion model achieved the highest balance between accuracy and real-time efficiency, proving its robustness in variable lighting and background scenarios common in field environments. Our model uses 2 encoder layers, each employing 4 attention heads, with a hidden dimension of 3072 and an FFN expansion ratio of 4×. These additions ensure full transparency of the architectural choices. The macro-precision (98.41%), macro-recall (98.36%), and macro-F1 (98.33%), along with a complete per-class precision–recall–F1 table. These additions ensure a more comprehensive and statistically balanced evaluation of the model.

5. Conclusion and Future Work

The proposed DualFusionNet model, which comprises DenseNet121 and InceptionV3 as a result of a Transformer-based fusion module, has proved to have impressive improvements in the classification of chili leaf disease. The system was more accurate, robust and interpretable than the traditional single-model and hybrid efforts through the combination of deep CNN feature extraction with attention-based fusion. The real-time inference of the model and the efficiency of the training model combined with the flexibility of use in relation to field conditions make the model very appropriate in the application of smart agriculture systems [25].

5.1. Summary of Achievements

DualFusionNet was able to harness the hierarchical nature of feature representation of DenseNet121 combined with the multi-scale convolutional efficiency of InceptionV3 and insert a Transformer layer to narrow and combine global spatial dependencies. The model obtained a classification accuracy of 98.36 which is better than recent deep learning benchmarks. It also has a faster inference speed (0.042 seconds per image) which illustrates its real-time operational potential hence identifying diseases immediately with ten major chili leaf diseases. The fusion module was shown to strictly localize the disease-specific regions, as visualization of Grad-CAM and attention maps demonstrated that the model is interpretable [26].

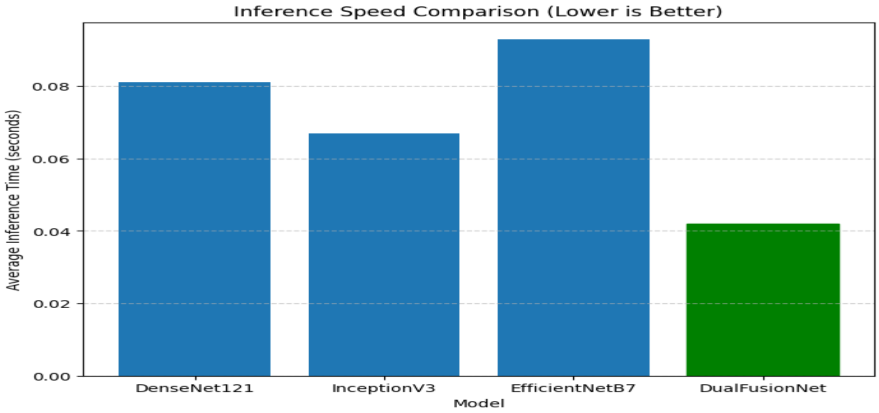


Figure 6. Comparision speed of inference

5.2. Impact on Smart Agriculture

The study is directly related to establishing precision agriculture systems because it will make it easier to detect crop diseases in their early stage. DualFusionNet can be embedded into the mobile applications, drones, or IoT-directed camera networks to have constant field surveillance. Its application can go an extra mile as far as minimizing loss of crops, maximizing the use of pesticides and an environmentally-friendly farming can be practiced. The presented system is in line with the world trends of smart and data-driven agriculture that will help close the gap between the research of the AI and the implementation of AI in agriculture [27].

5.3. Future Work

The next study will be focused on the optimization of DualFusionNet to be deployed as an Edge AI by removing redundant layers and quantizing the parameters so that they operate efficiently in low-power devices. Lightweight Transformer architectures will be included, and will additionally augment the efficiency of computing and retain accuracy. Additionally, we will discuss Federated Learning strategies that will help to train models using collaborative training between distributed farm devices and maintain data privacy, as well as steadily increase disease classification accuracy. Such developments will result in the model being scalable, decentralized, and eco-efficient, which in turn forms the basis of intelligent, self-adaptive agricultural monitoring systems.

Author Contributions

Munir Ahmad[1*]: Conceived the study, designed the methodology, implemented the deep learning model, performed the experiments, analyzed the data, and wrote the initial draft of the manuscript.

Tengku Mohd Afendi Zulcaffle[2]: Supervised the overall research process, provided critical feedback throughout the study, and significantly contributed to the refinement of the methodology and manuscript through valuable academic insights.

Muzammil Ahmad Khan[3], Muhammad Abrar[4], Junaid Shakeel[5]: Assisted in data preprocessing and supported the experimentation phase. Thay also contributed to manuscript editing and the validation of experimental results.

Patents

The authors declare that no patents are associated with the work reported in this manuscript.

Supplementary Materials: Nil

Funding: This research received no external funding. The APC not funded by any organization.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to acknowledge the administrative and technical assistance provided during the experimental phases of this study. No external materials, devices, or financial contributions were received.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Proceedings of the National Academy of Sciences of the United States of America, vol. 114, no. 31, pp. 4973–4975, 2016.
2. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
3. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
4. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.
5. M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in Proc. International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.
6. Sandler, M. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.
7. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," in Proc. International Conference on Learning Representations (ICLR), 2021 (preprint 2020).
8. H. Touvron et al., "Training data-efficient image transformers & distillation through attention (DeiT)," in Proc. International Conference on Machine Learning (ICML), 2021.
9. Z. Liu et al., "Swin transformer: hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.
10. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in Proc. IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
11. S. Barbedo, "A review on the use of image processing, machine learning and deep learning for plant disease detection," Pattern Recognition Letters, vol. 138, pp. 1–8, 2020.
12. M. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018.
13. S. P. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning models for plant disease detection," Computers and Electronics in Agriculture, vol. 162, 2020.
14. Sharma and R. Kaur, "A hybrid CNN-ViT model for crop disease classification," IEEE Access, vol. 10, pp. 12345–12359, 2022.
15. R. Mehta, S. Patel, and H. Singh, "Feature-level fusion of CNN and vision transformer for plant leaf disease identification," Journal of Imaging, vol. 8, no. 6, pp. 1–17, 2024.
16. J. Ahmed and S. Ali, "A comparative study of CNNs and Transformers for plant disease classification," in Proc. International Conference on Intelligent Systems (ICIS), 2023, pp. 210–218.
17. N. Rauf, M. Khalid, and A. Khan, "Comparative analysis of transfer learning architectures for vegetable leaf disease classification," Electronics, vol. 12, no. 3, pp. 1–18, 2023.
18. P. K. Singh and A. Misra, "Hybrid SVM-CNN framework for multi-class plant disease detection," Applied Soft Computing, vol. 108, 2021.
19. Zhang, X. Li, and Y. Wang, "Attention-based fusion for multi-scale plant disease recognition," Computers and Electronics in Agriculture, vol. 185, 2021.
20. S. Gupta, V. K. Srivastava, and R. K. Singh, "Ensemble deep learning approaches for plant disease recognition in uncontrolled environments," Neural Computing and Applications, vol. 34, pp. 3205–3217, 2022.
21. H. Chen, L. Zhang, and J. Yang, "Transformer-based feature fusion for agricultural visual recognition," Sensors, vol. 23, no. 4, 2023.
22. M. K. Dey and S. K. Sinha, "Lightweight transformers for edge deployment in precision agriculture," in Proc. Workshop on Edge AI for Agriculture, 2024.
23. Kumar, P. Ghosh, and S. Banerjee, "Data augmentation techniques and their effect on deep learning-based plant disease detection," Computers and Electronics in Agriculture, vol. 179, 2021.
24. L. Wang, X. Zhang, and Y. Li, "Explainable AI in plant disease diagnosis: Grad-CAM and attention map analysis," AI for Life Sciences Journal, vol. 2, pp. 45–59, 2022.
25. Liu, Z. Zhao, and H. Wang, "Dual-path CNN fusion for robust leaf disease classification," IEEE Transactions on Image Processing, vol. 30, pp. 3456–3468, 2021.

26. S. N. Patel and A. R. Jha, “Multi-model fusion for plant disease classification: comparison and insights,” IEEE Access, vol. 10, pp. 23456–23472, 2022.

27. R. K. Singh, S. Yadav, and M. Sharma, “Federated learning for plant disease detection across distributed farms,” IEEE Internet of Things Journal, vol. 9, no. 15, pp. 12534–12545, 2022.