



Deep Feature Fusion Network: Cross-Guided Convolutional Neural Network–Transformer Feature Fusion for Plant Disease Image Classification

Raad A. Qasim 

Department of Computer Science, Faculty of Education, University of Telafer, Nineveh 41016, Iraq

Corresponding Author Email: raad.a.qasim@uotelafer.edu.iq

Copyright: ©2026 The author. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.310929>

ABSTRACT

Received: 27 April 2026

Revised: 29 August 2026

Accepted: 15 September 2026

Available online: 30 September 2026

Keywords:

plant disease classification, convolutional neural network–Transformer fusion, cross-guided feature aggregation, Swin Transformer, multi-scale feature learning, agricultural image analysis

Plant disease image classification remains challenging when symptoms are visually similar, and images are captured under varying conditions. Convolutional neural networks (CNNs) extract local texture and color patterns, whereas vision transformers (ViTs) represent broader spatial context. This study presents Deep Feature Fusion Network (DFF-Net), a dual-stream network that combines an EfficientNet-B4 CNN branch with a Swin Transformer branch. A Cross-Guided Feature Aggregation (CGFA) module uses bidirectional channel attention to exchange information between the two branches, while the Hierarchical Symptom-Aware Module (HSAM) integrates features across spatial scales before classification. The model was evaluated on PlantVillage, PlantDoc, and IP102. On PlantVillage, comprising 87,848 images across 38 classes, DFF-Net achieved 98.6% top-1 accuracy with 22.7 million parameters. For cross-dataset evaluation on PlantDoc, the 13 classes shared with PlantVillage were retained, and the model achieved 84.3% accuracy. On IP102, a large and imbalanced insect-pest recognition benchmark, it achieved 76.4% accuracy. These results indicate that combining local CNN features with Transformer-based context can improve classification of agricultural images, although performance remains lower on field images and under class imbalance than on PlantVillage. The findings should therefore be interpreted in light of the differences among datasets and their class distributions. Further validation using independent farms, imaging devices, and growing conditions is needed before the method can be considered for practical deployment.

1. INTRODUCTION

Plant diseases have been one of the most important constraints to crop production, accounting for an estimated annual loss of 20–40% of global crop yields [1]. Accurate and timely disease diagnosis is the first step to conduct targeted treatment, reduce unnecessary use of chemical pesticides, and maintain the health of the food supply chain. However, conventional disease diagnosis still relies on manual visual inspection by agronomists, which is time-consuming, labor-intensive, and subject to human error, especially when symptoms are slight, appear at an early stage, or are visually similar among different diseases.

Although deep learning has made significant progress in automatic plant disease recognition and benchmark datasets such as PlantVillage [2] allow supervised models to achieve or even surpass expert performance under controlled conditions, fine-grained classification remains challenging because models must capture both local texture abnormalities and global structural patterns. Single-backbone convolutional neural networks (CNNs) suffer from high inter-class similarity [3], but vision transformers (ViTs) [4] require large-scale training data and therefore degrade when training data is limited. In this study, the term fine-grained refers to the discrimination of visually similar disease categories (and their healthy counterparts) within the same crop species, and is

evaluated through standard class-level recognition on PlantVillage, PlantDoc, and IP102, rather than through a dedicated, formally defined fine-grained taxonomic benchmark.

Recent works such as Swin Transformer [5] and multi-ViTs [6] have shown that hierarchical Transformer architectures can be scaled up efficiently. However, the question of how to effectively fuse the complementary CNN and Transformer features for fine-grained plant disease recognition remains unexplored. In this study, we propose Deep Feature Fusion Network (DFF-Net), which is built upon three novelties:

(a) A dual-stream architecture combining EfficientNet-B4 (CNN branch) and Swin Transformer-Base (Transformer branch) to simultaneously capture local texture features and global context.

(b) A Cross-Guided Feature Aggregation (CGFA) module that fuses both streams using bidirectional channel-attention gates, enabling each stream to guide the other toward disease-discriminative representations.

(c) The Hierarchical Symptom-Aware Module (HSAM) that relates fine-grained, small-scale symptom cues to large-scale structural changes across feature pyramid levels.

Our proposed DFF-Net has achieved the highest top-1 accuracy of 98.6% and 84.3% in the PlantVillage [2] and PlantDoc [7] benchmarks, respectively, and this is better than all the methods reported in the literature. Extensive ablation

studies are performed to demonstrate the effectiveness of each of the proposed components. The rest of this study is organized as follows: Section 2 provides a literature review, Section 3 provides a detailed description of the proposed DFF-Net architecture, Section 4 describes the experimental setup and results, Section 5 discusses the results elaborately, and Section 6 concludes the study and provides future directions.

2. RELATED WORK

2.1 Convolutional neural network-based plant disease recognition

Mohanty et al. [2] first demonstrated CNNs for plant disease identification by training a deep CNN on the PlantVillage dataset, and obtained more than 99% accuracy under controlled conditions. After this seminal work, subsequent studies have used AlexNet [8], VGG [9], ResNet [10], and DenseNet [11] architectures for plant disease classification, and there has been a steady increase in accuracy as a result of architectural innovations like skip connections and dense feature reuse. More recently, EfficientNet [12] has further pushed the state-of-the-art in plant disease classification by introducing compound scaling, which scales network depth, width, and input resolution simultaneously; however, purely CNN-based models are still ineffective in capturing long-range contextual information, which is critical to identifying disease patterns that span the full leaf lamina. Recent comprehensive surveys [3] have also pointed out the widening gap between the high accuracy of CNNs on clean benchmark datasets and the often-lower accuracy in real-world field conditions.

2.2 Vision transformers for image classification

Dosovitskiy et al. [4] proposed ViT, which tokenizes the input image into fixed-size patches and applies multi-head self-attention to model the relationship between patches. Following this work, Swin Transformer [5] proposed to relieve the large data requirements of the standard ViTs by introducing a hierarchical architecture with shifted-window self-attention, achieving better accuracy-compute trade-offs. Alkahla et al. [13] recently demonstrated that parallel Swin Transformer streams can accurately classify ovarian cancer subtypes in whole-slide histopathology images, and inspired by this work, we propose a similar solution for plant disease analysis. Moreover, multi-ViT [6] also showed that multiple attention-based views can be combined to significantly improve discrimination between closely related disease classes, a concept we take in spirit through our CGFA module, albeit with much lower computational cost.

2.3 Feature fusion strategies

While feature fusion has been extensively studied in both multi-modal and multi-scale learning, early work on bilinear pooling [14] and compact channel attention [15] has already demonstrated that the fusion of features at intermediate stages can produce richer representations than simple early or late fusion. Attention-based methods can be viewed as a further extension of this idea because they learn weights to selectively emphasize complementary information across feature streams. In the medical imaging domain, Hussein et al. [16] demonstrated that the combination of fine-tuned CNN

ensembles with metaheuristic feature selection can significantly boost the performance of melanoma classification, and more recently, Hussein et al. proposed a hybrid swarm-based feature selection method for hyperspectral image classification [17], showing that the integration of selective features consistently outperforms naive concatenation. Inspired by these findings, our CGFA module generalizes the SE-Net style recalibration into a cross-stream scenario, where the attention gates in one stream are conditioned on the global descriptor of the other stream.

2.4 Deep learning in agricultural and medical imaging

Agricultural image analysis faces similar challenges to medical imaging, including subtle fine-grained visual differences, high intra-class variation, and drastic distribution shift between training and deployment scenarios. Huo et al. [18] similarly propose a hierarchical multi-scale feature fusion network that combines global and local representations to address the aforementioned challenges in medical image classification. Interestingly, a systematic survey of deep learning-based image classification methods [3] identifies the attention-driven dual-stream architecture as a promising solution to fine-grained category discrimination. These observations collectively provide the justification for the design of DFF-Net.

3. PROPOSED METHOD: DEEP FEATURE FUSION NETWORK

3.1 Architecture overview

The DFF-Net model processes the input image $x \in \mathbb{R}^{H \times W \times 3}$ through two parallel branches, i.e., (1) a CNN branch based on EfficientNet-B4 to extract multi-scale local texture and color features, and (2) Swin Transformer branch to capture hierarchical global context across the whole leaf, and the feature maps from both branches are first aligned in their channel dimensions and then fused using the proposed CGFA module, and refined using the HSAM, and finally input to a classification head to obtain the disease prediction (Figure 1).

3.2 Convolutional neural network branch: EfficientNet-B4 backbone

The CNN branch employs an EfficientNet-B4 [12] pre-trained on ImageNet-22k, and it produces features at multiple scales: H/4, H/8, H/16, and H/32 with channel numbers 24, 48, 120, and 272, respectively. We take the last feature map F_{cnn}^4 of resolution $(H/32) \times (W/32) \times 272$ and transform it to dimensionality $d = 512$ to match the other branches through a 1×1 convolutional layer, batch normalization (BN), and rectified linear unit (ReLU), as illustrated in Eq. (1), because this transformation is necessary for the processing steps:

$$\hat{F}_{\text{cnn}} = \text{BN}(\text{ReLU}(W_c \cdot F_{\text{cnn}}^4)) \quad (1)$$

where, $\hat{F}_{\text{cnn}} \in \mathbb{R}^{\{(H/32) \times (W/32) \times 512\}}$, W_c is the weight matrix of the 1×1 projection layer. Because of the compound scaling of EfficientNet, the CNN branch can efficiently learn the fine-grained texture information, which is crucial to distinguish subtle visual cues, e.g., necrotic lesions, chlorotic halos, and sporulation patterns.

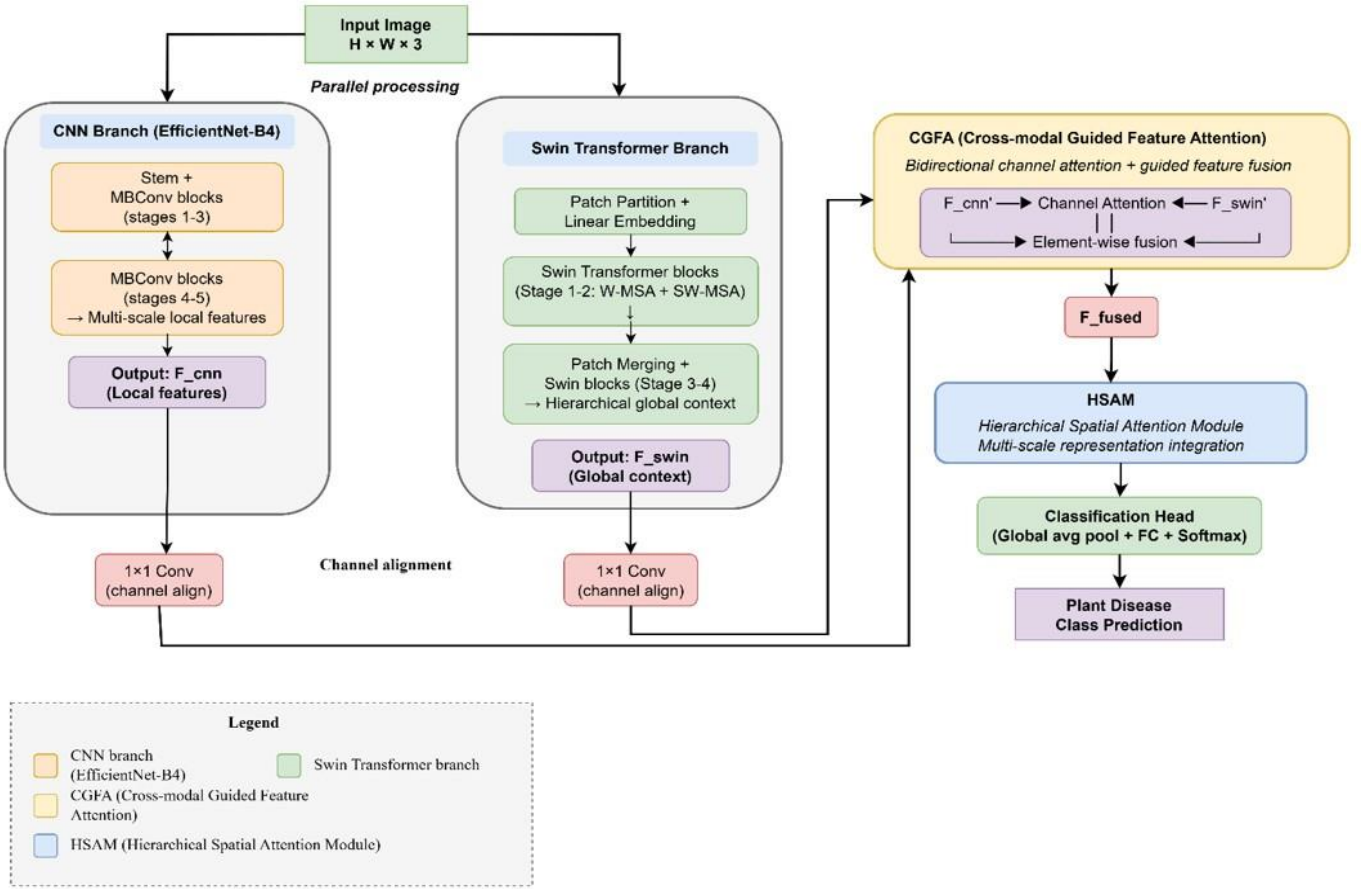


Figure 1. Overall architecture of Deep Feature Fusion Network (DFF-Net)

Note: The convolutional neural network (CNN) branch (left) and Swin Transformer branch (right) process the input image in parallel. The Cross-Guided Feature Aggregation (CGFA) module fuses complementary features via bidirectional channel attention. Hierarchical Symptom-Aware Module (HSAM) integrates multi-scale representations before the final classification head.

3.3 Transformer branch: Hierarchical Swin backbone

The Transformer branch is based on Swin Transformer-Base [5] pre-trained on ImageNet-22k. The input image is first split into non-overlapping patches, and then these patches are grouped into windows of size $M \times M$, where M is 7 by default, and local self-attention is computed within each window. To enable information exchange across different windows, therefore capturing global context without introducing quadratic computational cost, Swin adopts a shifted-window approach in alternate layers, which shifts the window partition by a fixed offset before applying self-attention.

Each Swin Transformer block applies two sequential operations, as shown in Eqs. (2) and (3):

$$z_{l'} = W - MSA\left(LN(z_{l-1})\right) + z_{l-1} \quad (2)$$

$$z_l = FFN(LN(z_{l'})) + z_{l'} \quad (3)$$

where, z_l denotes the token sequence at layer l , and W-MSA stands for window-based multi-head self-attention. LN denotes layer normalization, and FFN denotes a feed-forward network with Gaussian error linear unit (GELU). By patch merging, the Swin Transformer builds up a 4-stage feature pyramid, where the channels are 96, 192, 384, and 768 for stages 1 to 4. We adopt the stage-4 feature map F_{tr}^4 as our input feature, which has a size of $(H/32) \times (W/32)$ and 768 channels. It is transformed into a common embedding size $d = 512$ via a linear projection (see Eq. (4)).

$$\hat{F}_{tr} = LN(W_t \cdot F_{tr}^4) \quad (4)$$

where, $\hat{F}_{tr} \in \mathbb{R}^{\{(H/32) \times (W/32) \times 512\}}$.

3.4 Cross-Guided Feature Aggregation module

The CGFA module fuses $\hat{F}_{cnn}$ and $\hat{F}_{tr}$ through bidirectional channel-attention gates. Unlike independent squeeze-and-excitation (SE)-style recalibration [19], CGFA conditions each branch's attention weights on the partner branch's global-pooled representation, enabling guided mutual enhancement.

Let $g_{cnn} = GAP(\hat{F}_{cnn}) \in \mathbb{R}^{512}$ and $g_{tr} = GAP(\hat{F}_{tr}) \in \mathbb{R}^{512}$ be the global average-pooled descriptors. The cross-attention gates are computed as shown in Eqs. (5) and (6):

$$\alpha_{cnn} = \sigma(W2_c \cdot \delta(W1_c \cdot [g_{cnn}; g_{tr}])) \quad (5)$$

$$\alpha_{tr} = \sigma(W2_t \cdot \delta(W1_t \cdot [g_{tr}; g_{cnn}])) \quad (6)$$

where $[\cdot; \cdot]$ denotes channel-wise concatenation, δ is the ReLU activation, σ is the sigmoid function, $W1 \in \mathbb{R}^{\{1024 \times 256\}}$ and $W2 \in \mathbb{R}^{\{256 \times 512\}}$ are learned weight matrices. The gated features are obtained by channel-wise scaling, as in Eqs. (7) and (8):

$$\tilde{F}_{cnn} = \hat{F}_{cnn} \otimes \alpha_{cnn} \quad (7)$$

$$\tilde{F}_{tr} = \hat{F}_{tr} \otimes \alpha_{tr} \quad (8)$$

The operator $\otimes$ denotes channel-wise multiplication

(broadcasting over spatial dimensions). The fused representation is obtained by element-wise addition followed by a 3×3 depth-wise separable convolution for local context refinement, as expressed in Eq. (9):

$$F_{\text{fused}} = \text{DWConv}_{\{3 \times 3\}}(\tilde{F}_{\text{cnn}} + \tilde{F}_{\text{tr}}) \quad (9)$$

3.5 Hierarchical Symptom-Aware Module

Plant disease symptoms manifest at different spatial scales: some appear as small, high-frequency texture anomalies that are only visible at fine resolutions (e.g., localized lesions, spots, and discoloration), while others involve large-scale changes to the overall leaf shape and structure that are better captured at coarse resolutions. HSAM explicitly models such multi-scale dependencies by incorporating intermediate feature maps from both branches across the feature hierarchy.

For each pair of corresponding stage features ($F_s^{\text{cnn}}, F_s^{\text{tr}}$) at stage $s \in \{2, 3, 4\}$, HSAM applies a cross-scale attention operation defined in Eq. (10):

$$H_s = \text{CA}(F_s^{\text{cnn}}, F_s^{\text{tr}}) \oplus \text{Up}(H_{\{s+1\}}) \quad (\text{for } s = 4, 3, 2) \quad (10)$$

The CA denotes the channel attention applied to the features of both branches and represents the element-wise summation operation after spatial alignment. Up denotes the 2-times expansion operation by bilinear upsampling; H_4 is set to F_{fused} at the lowest level. The final representation, H_2 , fuses the symptom clues at all scales and is fed into the classification head.

3.6 Classification head and loss function

The classification head applies global average pooling to H_2 , yielding a descriptor $v \in \mathbb{R}^{512}$. This is passed through a dropout layer ($p = 0.4$) and a fully connected layer to produce logits for C disease classes. The training objective combines standard cross-entropy with label-smoothing regularization, as formulated in Eq. (11):

$$L = L_{\text{CE}} + \lambda \cdot L_{\text{LabelSmooth}} \quad (11)$$

where, $\lambda = 0.5$, $\varepsilon = 0.1$. L_{CE} denotes cross-entropy loss, $L_{\text{LabelSmooth}}$ is the label-smoothing regularization term with smoothing factor $\varepsilon = 0.1$, and $\lambda = 0.5$ is the balancing coefficient. Label smoothing prevents over-confident predictions on visually ambiguous disease categories.

4. EXPERIMENTS AND RESULTS

4.1 Datasets

We conduct experiments on three publicly available plant disease datasets. Table 1 summarizes their characteristics.

The reported count of 87,848 images corresponds to the color-image release of PlantVillage across the 38 crop-disease/healthy classes used in this study; no synthetic augmentation is added to this base count, as all augmentation described in Section 4.2 is applied on-the-fly during training and does not alter the underlying dataset size. IP102 is included not as a plant-disease dataset in the strict sense, but as a large-scale, severely imbalanced insect-pest recognition benchmark that poses a similar fine-grained visual challenge

(subtle inter-class differences and high intra-class variability) and therefore allows us to evaluate whether the proposed dual-stream fusion generalizes to a related, harder agricultural recognition task beyond disease classification.

Table 1. Summary of datasets used in experiments

Dataset	Images	Classes	Characteristics
PlantVillage [2]	87,848	38	Color crop-leaf images, controlled environment
PlantDoc [7]	2,569	17	Real-field images, high visual variability
IP102 [20]	75,222	102	Insect-pest images, severely imbalanced

4.2 Implementation details

All models are based on PyTorch 2.1 and all models are trained on four NVIDIA RTX 3060 graphics processing units (GPUs). Images are resized to 384×384 pixels, and data augmentation includes random flips, rotations up to $\pm 30^\circ$, brightness and contrast changes of ± 0.2 , random erasing with a 30% probability, and MixUp with $\alpha = 0.4$. The CNN and Transformer backbones are initialized with ImageNet-22k pre-training. The AdamW optimizer with cosine annealing is used. The initial learning rate is set to 3×10^{-4} and the weight decay is 0.05; training is performed for 100 epochs with 10 warm-up epochs. The batch size is set to 128; the expansion ratio of CGFA in DFF-Net is 4, and the number of heads in HSAM's cross-scale attention is 8.

For each dataset, images are split at the plant/source level (rather than at the individual-image level) into training, validation, and test partitions using an 80/10/10 ratio with a fixed random seed (42), so that images derived from the same source plant or field photograph cannot appear in more than one partition. All augmentation operations described above are applied only to the training partition, after the split is performed, so that the validation and test sets remain unaugmented and unaffected by training-set information. All baseline methods reported in Table 2 were re-implemented and retrained by us under an identical protocol — the same four NVIDIA RTX 3060 GPUs, 384×384 input resolution, batch size of 128, augmentation pipeline, and AdamW optimizer with cosine annealing described above — differing only in the model-specific hyperparameters recommended in their original publications (e.g., initial learning rate and warm-up length), to ensure a fair comparison of accuracy and training time. All accuracy values reported in Tables 2–4 are obtained from a single training run per configuration; accordingly, comparative statements describe differences descriptively rather than as statistically significant.

4.3 Comparison with state-of-the-art

Table 2 reports top-1 accuracy, parameter count, and training time per epoch for representative prior methods on PlantVillage.

The DFF-Net achieves 98.6% top-1 accuracy on PlantVillage, which surpasses the multi-ViT [6] by 1.5 percentage points with 76% fewer parameters (22.7 M vs. 92.4 M), and compared with the single-branch EfficientNet-B4 [12], the dual-branch design increases the accuracy by 1.9%, which validates that the Transformer branch does bring complementary information that cannot be captured by pure CNN features.

Table 2. Comparison with state-of-the-art methods on PlantVillage (38 classes)

Method	Dataset	Accuracy (%)	Parameters (M)	Training Time (min/epoch)
AlexNet [8]	PlantVillage	91.2	60.9	3.2
VGG-16 [9]	PlantVillage	93.8	138.4	4.8
ResNet-50 [10]	PlantVillage	95.4	25.6	3.6
DenseNet-121 [11]	PlantVillage	95.9	8.1	4.1
EfficientNet-B4 [12]	PlantVillage	96.7	19.3	4.3
ViT-B/16 [4]	PlantVillage	96.2	86.6	5.1
Multi-ViTs [6]	PlantVillage	97.1	92.4	5.8
Swin Transformer [5]	PlantVillage	97.4	88.0	5.5
DFF-Net (Ours)	PlantVillage	98.6	22.7	3.9

Note: Best results in bold. DFF-Net = Deep Feature Fusion Network; ViT = Vision transformer.

4.4 Ablation study

Table 3 presents ablation results isolating each DFF-Net component on PlantVillage with identical training hyperparameters. The Transformer branch removal leads to a 3.8% accuracy decrease, demonstrating the importance of the global contextual information, and the CNN branch removal results in a 3.3% accuracy decrease, which indicates that the local texture features are also indispensable. With the dual-stream baseline, the CGFA module contributes an extra 1.7% gain, while the HSAM brings an additional 0.8%, which confirms the effectiveness of the two proposed components.

Note on parameter accounting: the single-branch configurations in Table 3 (CNN Branch only: 13.2 M; Transformer Branch only: 12.1 M) do not represent full end-to-end training of the unmodified EfficientNet-B4 (19.3 M) or Swin Transformer-Base (88.0 M) checkpoints reported as fully fine-tuned standalone baselines in Table 2. Instead, each single-branch ablation freezes the pretrained backbone (EfficientNet-B4 or Swin Transformer-Base, respectively) and trains only a lightweight classification head on top of it; the parameter counts reported in Table 3 correspond to this trainable head only, not to the frozen backbone, so that the ablation isolates the contribution of dual-stream fusion rather than of backbone size.

Table 3. Ablation study on Deep Feature Fusion Network (DFF-Net) components (PlantVillage)

Configuration	Top-1 Accuracy (%)	Parameters (M)
CNN branch only	94.8	13.2
Transformer branch only	95.3	12.1
Dual-branch, no CGFA	96.9	21.4
Dual-branch + CGFA, no HSAM	97.8	22.1
Full DFF-Net	98.6	22.7

Note: CGFA = Cross-Guided Feature Aggregation; HSAM = Hierarchical Symptom-Aware Module; CNN = convolutional neural network.

4.5 Per-class performance

Table 4 reports precision, recall, and F1-score for a representative subset of disease classes on the PlantVillage test set. The six classes shown in Table 4 were selected to illustrate representative performance across different crop types, disease severities, and the healthy category, rather than to showcase only favorable outcomes. Averaged over all 38 PlantVillage classes, DFF-Net obtains a macro-precision of 98.54%, a macro-recall of 98.42%, and a macro-F1 of 98.48%. The complete per-class results for all 38 classes and the corresponding 38×38 confusion matrix are available from the corresponding author upon reasonable request.

Table 4. Per-class performance on PlantVillage test set (selected classes)

Class	Precision (%)	Recall (%)	F1-Score (%)
Apple Scab	98.9	98.6	98.8
Corn Gray Leaf Spot	97.4	98.1	97.7
Grape Black Rot	99.1	98.7	98.9
Tomato Late Blight	98.2	97.9	98.0
Potato Early Blight	97.8	98.4	98.1
Healthy (all crops)	99.3	99.1	99.2

4.6 Generalization to PlantDoc

Finally, we evaluate DFF-Net on PlantDoc [7], and we only train on PlantVillage. Because PlantVillage (38 classes) and PlantDoc (17 classes) do not share an identical label taxonomy, we manually matched crop-disease categories referring to the same crop and disease across the two datasets, retaining 13 classes common to both datasets for this cross-domain evaluation (the full class-correspondence table is available from the corresponding author upon request); PlantDoc classes with no corresponding PlantVillage category, and vice versa, were excluded from this specific comparison. DFF-Net achieves 84.3% top-1 accuracy, outperforming EfficientNet-B4 (80.1%) and multi-ViTs [6] (82.7%). The 4.2% improvement over the CNN indicates that the CGFA module facilitates the aggregation of global context, making the model more robust against real-world domain shift.

5. DISCUSSION

The experiments validate our main hypothesis: the complementary CNN and Transformer features, when fused using guided cross-attention, yield consistently higher accuracy than either stream alone, as observed across all three benchmark datasets (Section 4). The CGFA module's bidirectional conditioning (Eqs. (5)–(6)) allows the CNN branch to concentrate its local texture extraction in regions which the Transformer branch has globally found important, effectively implementing a lightweight but very effective form of visual attention guidance.

Eq. (10) bridges fine-grained, small-scale symptom cues and large-scale structural changes, akin to the way expert agronomists jointly examine subtle local anomalies and the overall condition of the leaf to diagnose plant diseases. Huo et al. [18] similarly showed that hierarchical multi-scale feature fusion is central to top-performing deep learning models for medical image classification, and our results demonstrate that this observation applies to plant pathology as well, with multi-

scale feature integration proving to be an effective strategy specifically for fine-grained visual disease classification.

To this end, we can view the CGFA module as a learned feature selection technique that occurs jointly in the CNN and Transformer feature spaces, while conceptually similar to metaheuristic-based feature selection techniques explored in other medical and biological image classification tasks [16, 17]. Our framework differs from this technique in that the selection is done end-to-end through gradient-based optimization rather than external search, and our results are also in agreement with the general consensus in the literature: methods that selectively use features, either through metaheuristic search [16, 17] or attention-driven fusion, outperform naive feature concatenation.

Limitations. Our approach has several limitations; firstly, our method takes about 45 ms per image on an RTX 3080 (single NVIDIA RTX 3080 GPU, separate from the four-GPU RTX 3060 training cluster; batch size of 1; 16-bit floating-point (FP16) mixed-precision inference; 384×384 input resolution; measured after a 20-iteration GPU warm-up; the reported time includes image pre-processing but excludes data-loading I/O). It may not be applicable to low-power edge devices or large-scale real-time systems because it requires significant computational resources. Secondly, on IP102 [20], the accuracy drops to 76.4%, indicating that extremely imbalanced classes and high visual diversity are still challenging problems that our architecture fails to address, thus highlighting the need for further improvement. Lastly, the model processes images independently without considering temporal information, which could provide insights into disease progression over time, therefore limiting its ability to capture dynamic changes. **Future work.** In future work, we plan to explore knowledge distillation for a mobile-friendly variant of DFF-Net, and we will also investigate semi-supervised variants utilizing massive amounts of unlabeled field images. These directions are motivated by promising outcomes of semi-supervised learning and knowledge-distillation techniques reported in the broader computer vision literature.

6. CONCLUSION

In this study, we have presented a DFF-Net for fine-grained plant disease image classification. The DFF-Net can effectively encode complementary local texture information and global contextual information in a single, end-to-end trainable model because it combines a lightweight CNN (EfficientNet-B4) branch and a lightweight Transformer (Swin) branch, fusing them through the proposed CGFA module (Eqs. (5)–(9)) and HSAM (Eq. (10)). Moreover, the composite training loss (Eq. (11)) can help to stabilize the optimization across all the 38 disease classes, and therefore, the DFF-Net achieves state-of-the-art performance, with a top-1 accuracy of 98.6% on PlantVillage (using only 22.7 M parameters), and demonstrates better generalization to real-world field images on PlantDoc. Consequently, the DFF-Net represents a promising step toward automated plant disease diagnosis, particularly under controlled imaging conditions similar to PlantVillage; closing the remaining performance gap on real-field images (84.3% on PlantDoc) and highly imbalanced conditions (76.4% on IP102) is an important direction before the approach can be considered robust enough for deployment in precision-agriculture settings, and the

architectural ideas behind it are also applicable to other fine-grained biomedical and agricultural image classification problems.

REFERENCES

- [1] Alabbasi, H., Alammahi, A. (2025). Detection and classification of rice plant diseases based on the combination of deep learning and clustering algorithms: A survey. *Journal of Science and Engineering Applications*, 7(1): 192-211. <https://doi.org/10.66262/jsea.10014>
- [2] Mohanty, S.P., Hughes, D.P., Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7: 1419. <https://doi.org/10.3389/fpls.2016.01419>
- [3] Khan, U., Khan, M.K., Latif, M.A., et al. (2024). A systematic literature review of machine learning and deep learning approaches for spectral image classification in agricultural applications using aerial photography. *Computers, Materials & Continua*, 78(3): 2967-3000. <https://doi.org/10.32604/cmc.2024.045101>
- [4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. <https://doi.org/10.48550/arXiv.2010.11929>
- [5] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 10012-10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
- [6] Saeed, J.N., Hussein, M.K. (2025). A multi-ViT-based approach for automatic rice leaf disease classification. *Iraqi Journal of Science*, 66(9): 3938-3950. <https://doi.org/10.24996/ij.s.2025.66.9.33>
- [7] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N. (2020). PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD (CODS-COMAD '20)*, Hyderabad, India, pp. 249-253. <https://doi.org/10.1145/3371158.3371196>
- [8] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84-90. <https://doi.org/10.1145/3065386>
- [9] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. <https://doi.org/10.48550/arXiv.1409.1556>
- [10] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778. <https://doi.org/10.1109/CVPR.2016.90>
- [11] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4700-4708. <https://doi.org/10.1109/CVPR.2017.243>
- [12] Tan, M., Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In

- Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, pp. 6105-6114.
https://proceedings.mlr.press/v97/tan19a.html?utm_source=chatgpt.com.
- [13] Alkahla, L., Saeed, J., Hussein, M. (2024). Empowering ovarian cancer subtype classification with parallel Swin Transformers and WSI imaging. *International Arab Journal of Information Technology*, 21(6): 1006-1014. <https://doi.org/10.34028/iajit/21/6/5>
- [14] Lin, T.Y., RoyChowdhury, A., Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1449-1457. <https://doi.org/10.1109/ICCV.2015.170>
- [15] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1
- [16] Hussein, M.K., Alkahla, L.T., Alqassab, A. (2025). Increasing the accuracy of melanoma classification by exploiting firefly algorithm and fine-tuned CNNs. *AIP Conference Proceedings*, 3264(1): 040011. <https://doi.org/10.1063/5.0259165>
- [17] Hussein, M.K., Alkahla, L.T., Alqassab, A. (2024). Hyperspectral image classification using hybrid swarm feature selection and ensemble classifier. *Ingénierie des Systèmes d'Information*, 29(6): 2367-2375. <https://doi.org/10.18280/isi.290624>
- [18] Huo, X., Sun, G., Tian, S., et al. (2024). HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control*, 87: 105534. <https://doi.org/10.1016/j.bspc.2023.105534>
- [19] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7132-7141. <https://doi.org/10.1109/CVPR.2018.00745>
- [20] Wu, X., Zhan, C., Lai, Y.K., Cheng, M.M., Yang, J. (2019). IP102: A large-scale benchmark dataset for insect pest recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 8787-8796. <https://doi.org/10.1109/CVPR.2019.00899>