DenseSwinLight: A Hybrid CNN–Transformer Model with Lightweight Post-hoc Fusion for Visual Explainability

Deric Claudio Vitasoa
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
dericclaudiovitasoa@gmail.com

Paul Mahenina Randriamitsiry
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
mrandriamitsiry@gmail.com

Hajarisena Razafimahatratra
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
hajarisena@yahoo.fr

Thomas Mahatody
Laboratory for Mathematical and Computer Applied to the Development Systems (LIMAD)
University of Fianarantsoa
Fianarantsoa, Madagascar
hasinathomas@hotmail.com

***Abstract*— The Hybrid CNN–Transformer architectures offer high-performance image analysis by combining local extraction and global contextual modeling. However, this integration complicates decision-making mechanisms and increases the opacity of models, making them difficult for human users to interpret. We propose DenseSwinLight, a hybrid approach that combines convolution-based feature extraction using DenseNet201 and global contextual modeling using the Swin Transformer V2-Large to simultaneously enhance local discrimination and global context. Beyond performance, the goal of this work is to make the model's decisions more transparent. We introduce a post-hoc explainability module based on the fusion of two complementary sources of evidence, namely a Grad-CAM map from the CNN branch and a proxy attention map. The fusion is learned by an extremely lightweight MLP fusion module, consisting of only 65 parameters, and constrained by area, total variation, and entropy regularizations to produce a parsimonious and stable explanatory mask. Evaluated on a real dataset acquired in an uncontrolled environment covering six classes (Bacteria, Fungi, Healthy, Pest, Phytophthora, and Virus), DenseSwinLight achieves an accuracy of 98.96% on the test set with Test-Time Augmentation, and a macro F1 score of 0.9897. The fusion module achieves a combined soft fidelity score of 0.498 and 0.513 in hard mode according to the insertion and deletion curves, for an average mask area of 0.12. These results confirm the model's ability to provide robust predictions while producing compact and actionable visual explanations for human interpretation.**

Keywords—CNN–Hybrid Transformer; DenseNet201; Swin Transformer V2; Grad-CAM; Heatmap fusion; Explainable AI; Insertion–Deletion; Precision agriculture; Leaf diseases.

I. INTRODUCTION

Deep learning has revolutionized image analysis, enabling the development of automated systems capable of making decisions in complex environments. However, the predictive power of deep models comes with increasing opacity, making it difficult for human users to understand the decision-making mechanisms. This issue, often referred to as the "black box" problem, limits the adoption of AI in sensitive areas where transparency and trust are essential [1,2,3,4].

To address this challenge, research into XAI (eXplainable Artificial Intelligence) has intensified, aiming to make deep learning models more interpretable and provide understandable justifications for their predictions. Explainability approaches generally fall into two categories: post-hoc methods, which generate explanations after the model has been trained, and intrinsically interpretable architectures, designed to produce transparent decisions from the outset [1-4].

In the context of hybrid models combining convolutional neural networks (CNNs) and Transformers, the issue of explainability becomes particularly crucial. While CNNs are renowned for their ability to extract local features, Transformers provide global modeling thanks to their attention mechanisms. However, the fusion of these two paradigms further complicates the understanding of decisions, hence the need to develop explainability methodologies adapted to these hybrid architectures [2,3,5].

Recent work shows that multi-source visual explainability, based on a combination of local and global activation maps, makes it possible to better interpret the decisions of hybrid models and increase user confidence. Furthermore, quantitative and human evaluation of the quality of explanations is becoming a key issue in ensuring the relevance and usefulness of AI systems in practice [4,5,8,29].

This work provides two main contributions: a robust hybrid classification module combining DenseNet201 and Swin Transformer V2-Large, integrating advanced regularization strategies and inference by Test-Time Augmentation, and a post-hoc explainability module based on the learned fusion of Grad-CAM maps and proxy attention, producing compact explanatory masks evaluated by insertion and deletion metrics [2-5].

The structure of the article is as follows. Section II reviews related work on CNN–Transformer hybrid models and explainability methods in computer vision. Section III presents the proposed DenseSwinLight methodology, including the hybrid architecture, explainability mechanisms, fusion module, and processing pipeline. Section IV describes the experimental validation framework and reports the obtained results. Section V discusses the limitations and future research directions. Finally, Section VI concludes the paper.

II. Related Work

Explainable Artificial Intelligence (XAI) has emerged as a major area of research, particularly in the field of deep learning, where understanding so-called "black box" models remains a key challenge. Several post-hoc approaches, such as LIME, SHAP, Grad-CAM, and saliency maps, have been widely adopted to provide local and visual explanations of convolutional neural network decisions. Recent synthesis work has proposed comprehensive overviews of these techniques, addressing their methodological foundations, associated evaluation metrics, and open challenges related to the robustness and fidelity of explanations [7,8].

Recent advances in computer vision have introduced hybrid CNN–Transformer architectures that combine local feature extraction with global contextual modeling through attention mechanisms. However, interpreting these hybrid models remains challenging due to possible misalignment between convolutional activations and Transformer attention regions. Several studies have addressed this issue through interpretable architectures and explanation fusion strategies, particularly in medical applications [2,3,5,8].

Existing explainability methods for CNNs and Transformers exhibit complementary strengths and weaknesses in terms of localization precision, contextual understanding, robustness, and fusion stability, as summarized in Table I.

TABLE I. EXPLAINABILITY FUSION APPROACHES

Method	Fusion Type	Explainability	Limitation
Grad-CAM	CNN only	Local	No global context [27]
Attention	Transformer only	Global	Low localization precision [8,15]
Avg Fusion	Static	Weak adaptive capability	Sensitive to noisy activations
Product Fusion	Static	Strong activation amplification	Unstable and fragmented maps
FusionMLP (Proposed)	**Learnable**	**Local + Global**	**Lightweight adaptive fusion**

CNN-based explainability approaches such as Grad-CAM provide accurate lesion localization but lack explicit modeling of long-range contextual dependencies [27]. Conversely, Transformer attention mechanisms capture broader contextual interactions but may generate spatially diffuse explanations with limited localization precision [5,8,15]. Existing heuristic fusion approaches, including averaging and element-wise multiplication, remain static and sensitive to noise amplification. These limitations motivate the development of the proposed FusionMLP, which introduces a lightweight learnable fusion strategy capable of adaptively combining local and global explanatory evidence while preserving computational efficiency.

Recent hybrid CNN–Transformer approaches have demonstrated the effectiveness of combining local feature extraction and global contextual modeling for medical and agricultural image analysis [2,3,5,28,31].

Despite recent advances, existing explainability approaches remain limited by high computational cost, isolated CNN or Transformer interpretations, and heuristic fusion strategies rarely validated with fidelity metrics. To address these issues, DenseSwinLight combines a hybrid CNN–Transformer architecture with a lightweight post-hoc fusion module for generating compact and reliable visual explanations [30].

III. Methodology

In this chapter, we present the architecture of our DenseSwinLight hybrid model and describe the local and global explainability mechanisms generated. We present in detail the design of the post-hoc fusion module, whose particularity is to combine heterogeneous sources of explanation in a manner that is interpretable and faithful to the model's decision-making process.

A. Architectural composition of DenseSwinLight

This section describes the modular architecture of DenseSwinLight illustrated in Fig. 1, highlighting the hybrid classification block and the explainability fusion chain used to generate fused interpretable evidence maps.

- **DenseSwinNet block:** The predictive core of DenseSwinLight is based on the DenseSwinNet block [32], a sequential architecture combining DenseNet201 and Swin Transformer V2. DenseNet201 extracts discriminative local features from the input image $x \in \mathbb{R}^{H\times W\times 3}$, producing deep maps $F_{CNN} \in \mathbb{R}^{h\times w\times c}$. To ensure compatibility with the Transformer, these representations are projected by a convolution followed by bilinear interpolation to align the spatial dimensions and number of channels.

$$F_{proj} = \text{Interp}(\text{Conv}_{1x1}(\text{F}_{\text{CNN}})) \quad (1)$$

The projected representation is then partitioned into patches and transformed into tokens via a linear embedding, which constitutes the input to the Swin Transformer V2:

$$Z_0 = \text{Embed}(\text{Patch}(\text{F}_{\text{proj}})) \quad (2)$$

the Swin Transformer V2 processes these tokens through several hierarchical stages using shifted windows attention. This strategy makes it possible to capture both local and global dependencies while controlling computational complexity. Finally, an MLP

classification head generates the probability distribution over pathological classes [19].

- **Attention Map Block:** The proxy attention module extracts global contextual evidence from the Transformer branch using a gradient-based Grad×Input mechanism applied to Transformer feature activations. The resulting proxy attention map is normalized to produce a compact and interpretable global explanation.
- **Grad-CAM block:** At the same time, DenseSwinLight generates a Grad-CAM map on the CNN side in order to locate the most discriminating areas for the predicted class. For a class c, the Grad-CAM map is defined by:

$$H_{gc} = ReLU(\sum_K \alpha_K^C F_K), \alpha_K^C = \frac{1}{Z}\sum_{I,J} \frac{\partial Y^C}{\partial F_K^{IJ}} \quad (3)$$

this formulation is based on the gradients of the output associated with the target class relative to convolutional activations. It provides accurate local evidence, complementary to the global attention of the Transformer [27].

- **Post-hoc Interpretable Fusion Block**: The originality of DenseSwinLight lies in its interpretable post-hoc fusion module, designed to combine two heterogeneous explanations: the Grad-CAM H_{gc} heatmap (local evidence) and the H_{att} attention heatmap (global evidence). Before fusion, the two maps are spatially aligned and normalized to reduce scale bias. Fusion is then performed using a lightweight multilayer perceptron f_θ, applied to the concatenation of the two maps :

$$H_{fuse} = f_\theta([H_{gc}, H_{att}]) \quad (4)$$

unlike heuristic fusions, this mechanism learns an adaptive weighting of local and global contributions, producing a more stable unified map H_{fuse} that is better aligned with the final decision. The fusion module is optimized under the constraint of the following objective function :

$$\mathcal{L}_{fusion} = \lambda_1 \mathcal{L}_{area} + \lambda_2 \mathcal{L}_{TV} + \lambda_3 \mathcal{L}_{entropy} \quad (5)$$

where λ_1 , λ_2 and λ_3 are weighting coefficients controlling the importance of area sparsity, spatial smoothness, and entropy regularization, respectively. $\mathcal{L}_{area}$ promotes spatial parsimony of the explanatory mask, $\mathcal{L}_{TV}$ imposes local regularity to reduce noise, and $\mathcal{L}_{entropy}$ encourages an informative distribution of activations.

The proposed FusionMLP was designed with only 65 trainable parameters to minimize computational overhead while preserving adaptive fusion capability. Unlike heuristic fusion methods such as averaging or element-wise multiplication, the lightweight MLP dynamically learns the contribution of local and global evidence, producing more stable and faithful explanations.

B. Steps for implementing the approach

This section presents the operational pipeline of DenseSwinLight, including data preparation, dataset partitioning, hybrid classification, and fusion of explainability maps for unified visual interpretation.

Step I-II: Data Preparation and Preprocessing Pipeline. Raw RGB leaf images are resized to 224×224, normalized, and augmented using geometric and photometric transformations to simulate uncontrolled acquisition conditions, including horizontal flipping, rotation (±20°), scaling (0.8–1.2), gamma correction (0.7–1.5), Gaussian noise injection (σ = 0.05), and PCA color jitter (α = 0.1). Data augmentation is applied only during training, whereas validation and test sets use only resizing and normalization to ensure unbiased evaluation and reliable explainability analysis [16].

Step III: Data Splitting. After preprocessing, the dataset is partitioned into three disjoint subsets to ensure reliable evaluation: 70% for training, 20% for validation, and 10% for final testing. The distribution is stratified to preserve the proportions of classes in each subset, which limits sampling bias and stabilizes metrics [22].

Step IV: Hybrid Classification Module. The preprocessed images are first encoded by a CNN backbone to extract local discriminative features, then projected and partitioned into patches before being processed by a hierarchical Transformer to capture global contextual dependencies through attention. Finally, a classification head predicts the probability of each pathological class. This CNN–Transformer combination improves robustness under uncontrolled conditions [19,26].

Step V: Heatmap Fusion Mechanism. To obtain more stable explainability, we extract two complementary maps: Grad-CAM (CNN-oriented local evidence) and proxy attention map (Transformer-oriented global evidence). These heatmaps are then resized and normalized in the same space, then combined via a lightweight MLP-type post-hoc fusion module, which learns to weight their contributions in order to produce a unified evidence map aligned with the model's final decision.

IV. VALIDATION OF THE STRATEGY

This section presents the empirical evaluation of DenseSwinLight using a unified potato leaf disease dataset built from three complementary collections acquired under real-world conditions. It also describes the dataset partitioning strategy and analyzes the obtained performance and explainability results.

A. Case Study

a) Dataset : DenseSwinLight is evaluated on a six-class potato leaf disease classification task (Bacteria, Fungi, Healthy, Pest, Phytophthora, Virus) using a combined dataset built from

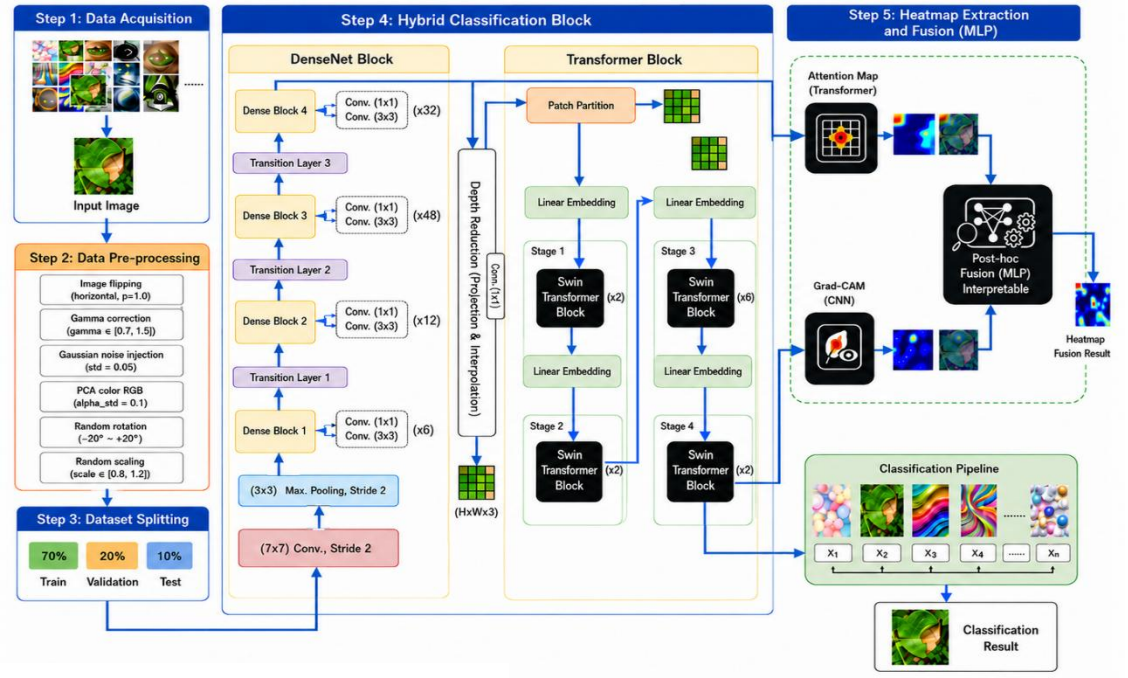

Fig. 1. Architectural diagram of the proposed DenseSwinLight model.

PLD-UE [21], PlantVillage [20], and the PLD dataset to increase intra-class variability and real-world diversity. The class distribution is presented in Table II.

TABLE II. DATASET DISTRIBUTION.

Class	PLD-UE	PlantVillage	PLD	Total
Bacteria	569	-	-	569
Fungi	748	1000	1628	3376
Healthy	201	152	1020	1373
Pest	611	-	-	611
Phytophthora	347	1000	1424	2771
Virus	532	-	-	532
Total	**3008**	**2152**	**4072**	**9232**

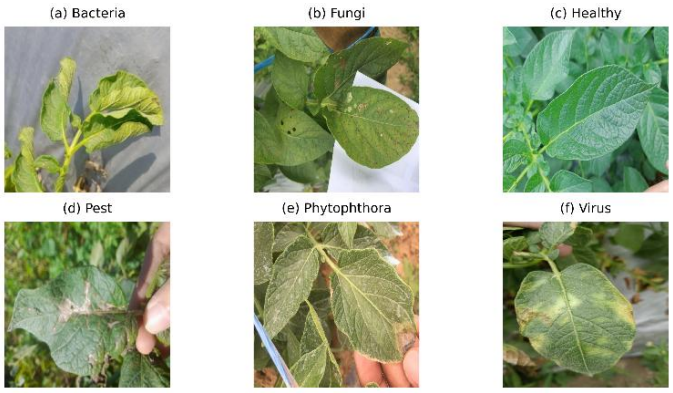

Fig. 2. Representative samples from the dataset illustrating the six considered classes: (a) Bacteria, (b) Fungi, (c) Healthy, (d) Pest, (e) Phytophthora, and (f) Virus.

b) Dataset merging strategy : To improve the robustness and generalization capability of DenseSwinLight, three complementary datasets were merged: PlantVillage, PLD, and PLD-UE. PlantVillage mainly contains images acquired in controlled environments, whereas PLD and PLD-UE introduce real-world variability related to lighting, background complexity, and field acquisition conditions. As illustrated in Fig. 3, the fusion of these datasets increases data diversity and improves the model's robustness under uncontrolled conditions.

c) Data augmentation strategy and partitioning protocol : For the training of DenseSwinLight, a data preparation strategy was implemented to improve the robustness and generalization of the model. Targeted data augmentation was applied to reduce inter-class imbalance, leading to a final balanced distribution of 3,376 images per class. The transformations selected, detailed in Table III, model common disturbances in real-world conditions, such as variations in orientation, scale, illumination, and noise, thereby helping to strengthen the model's generalization ability.

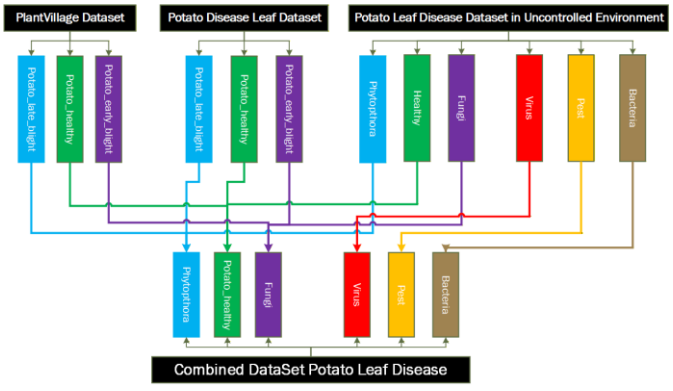

Fig. 3. Data Set Merging Process.

TABLE III. DATA AUGMENTATION.

Technique	Objective	Parameters
Horizontal flipping	Orientation invariance	p = 1.0
Gamma correction	Illumination variation	$\gamma \in [0.7, 1.5]$
Gaussian noise	Noise robustness	$\sigma = 0.05$
PCA color (RGB)	Color variation	$\alpha = 0.1$
Random rotation	Rotation invariance	±20°
Random scaling	Scale robustness	[0.8, 1.2]

After data augmentation, the dataset of 20,256 images was split into training, validation, and test sets following a 70%/20%/10% distribution [16]. Each class contains 2,363 training images, 675 validation images, and 338 test images, ensuring reliable optimization and unbiased generalization evaluation.

B. *Implementation details*

This section describes the main implementation choices, including the runtime environment, hyperparameters, and optimization strategies selected for training the model.

a) Model hyperparameters: Training was performed progressively using frozen and partially unfrozen stages with AdamW optimization [23], cosine LR scheduling, AMP acceleration, and TTA-based inference.

TABLE IV. HYPERPARAMETERS.

Category	Hyperparameters
Input & batch	Input size: 224 × 224; training batch size: 32 (Phases 1 & 2.1); VRAM-safe batch size: 8 with gradient accumulation ×4 (effective batch = 32)
Optimization	Optimizer: AdamW; LR Phase 1: 3e−4; LR Phase 2.1: 1e−5; LR Phase 2.2: 8e−7; weight decay adapted per phase
Regularization & augmentation	Dropout = 0.5 in the classifier; CutMix p = 0.5
Acceleration	AMP (torch.amp); gradient checkpointing

b) Hardware and software environment: The experiments were performed on a Linux platform equipped with an NVIDIA L4 GPU with 22.16 GB of video memory. Training and inference were accelerated by CUDA and cuDNN v9, using PyTorch 2.9 with automatic mixed precision (AMP) and gradient checkpointing to optimize memory usage. All experiments were conducted in a Google Colab environment, ensuring the reproducibility of the results.

TABLE V. HARDWARE ENVIRONMENT.

Category	Specification
Platform	Linux (Google Colab)
Programming language	Python 3.12.12
GPU	NVIDIA L4 (22.16 GB VRAM)
Driver / CUDA	NVIDIA Driver 550.54.15; CUDA 12.6
Deep learning framework	PyTorch 2.9.0; Torchvision 0.24.0
Pre-trained models	timm 1.0.24

To further characterize the efficiency of DenseSwinLight, additional computational complexity metrics were analyzed, including parameter count, memory consumption, inference speed, and computational cost.

TABLE VI. COMPUTATIONAL COMPLEXITY

Metric	Value
FusionMLP Params	65
Total Model Params	≈ 224 M
GPU Memory	≈ 13.8 GB
Inference Time	≈ 42 ms/image
FLOPS	≈ 43.7 GFLOPS

All experiments were conducted using fixed random seeds for NumPy, PyTorch, and CUDA to ensure reproducibility. Each experiment was repeated three times and the reported results correspond to the mean performance.

C. *Evaluation metrics*

This section presents the metrics used to evaluate the classification performance and the quality of the explanations generated by the proposed model.

a) Classification performance : DenseSwinLight achieves 98.96% test accuracy (TTA), with a macro precision of 0.9897, a macro recall of 0.9896, and a macro F1 score of 0.9897, demonstrating strong generalization under uncontrolled conditions. The accuracy and loss curves indicate stable convergence throughout training.

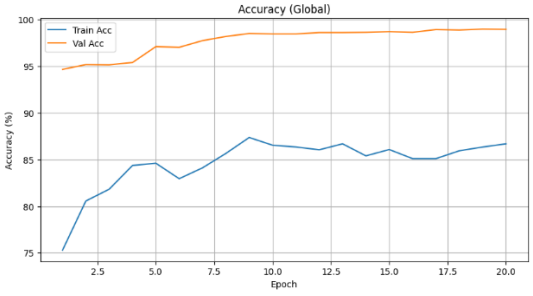

Fig. 4. Evolution of accuracy (train/val) during overall training.

The rapid increase in validation accuracy reflects efficient optimization and stable convergence behavior throughout the training process, suggesting that the proposed model generalizes effectively to previously unseen samples acquired under real-world and uncontrolled environmental conditions, while

maintaining strong classification consistency across different pathological classes.

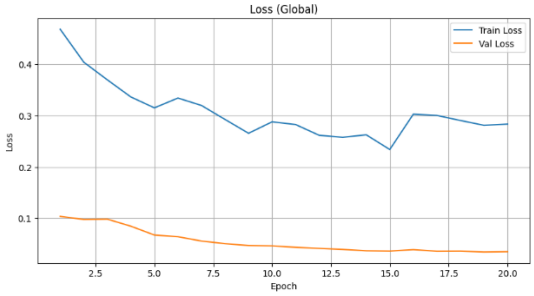

Fig. 5. Evolution of loss (train/val) during training.

The sustained decrease in loss validation suggests stable optimization under regularization.

b) Analysis by class and confusion matrix : The confusion matrix (Fig. 6) shows near-perfect recognition of the Bacteria class, with only minor confusion between visually similar classes under uncontrolled conditions. The classification report confirms high and balanced F1 scores across all categories, demonstrating the robustness of the proposed CNN–Transformer architecture.

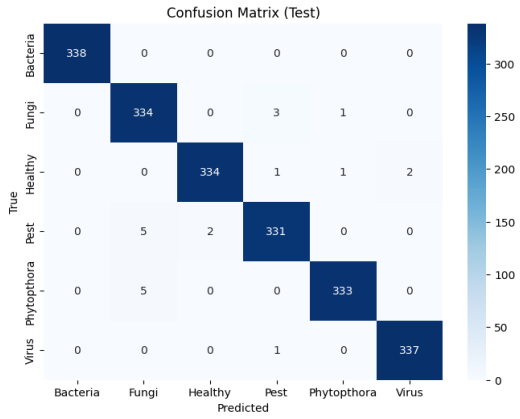

Fig. 6. Confusion matrix on the test set.

c) Multiclass ROC analysis : The ROC analysis confirms excellent class separability, with AUC values consistently close to 1.0 across all classes.

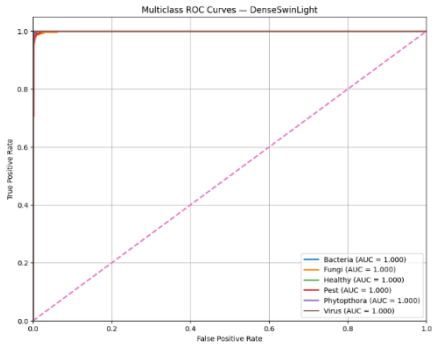

Fig. 7. Multiclass ROC curves obtained for the six potato leaf diseases.

TABLE VII. ROC-AUC SCORES

Class	AUC
Bacteria	1.0000
Fungi	0.9995
Healthy	0.9999
Pest	0.9996
Phytophthora	0.9999
Virus	1.0000

d) Explainability results : The MLP fusion module, trained under explicit area constraints to combine Grad-CAM and proxy attention maps, shows stable optimization behavior with loss values around 0.66–0.67 (Fig. 8), consistent with the multi-term objective function integrating fidelity, total variation, and entropy regularization. The fidelity curves shown in Fig. 9 exhibit high soft insertion performance (AUC = 0.7285) and lower soft deletion scores (AUC = 0.2681), which is characteristic of conservative and compact explanatory masks. The hard mode provides more balanced fidelity behavior (0.513), confirming the stability and relevance of the proposed fusion strategy.

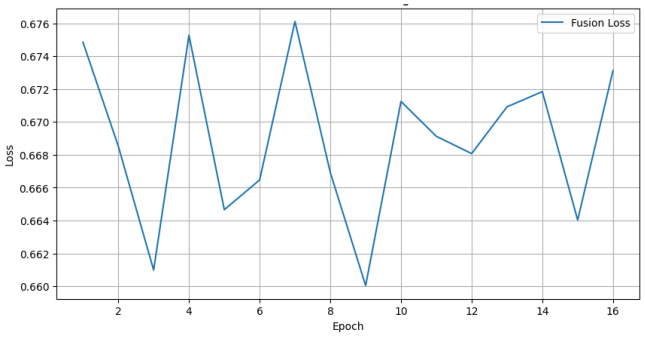

Fig. 8. Training loss curve of the MLP fusion module.

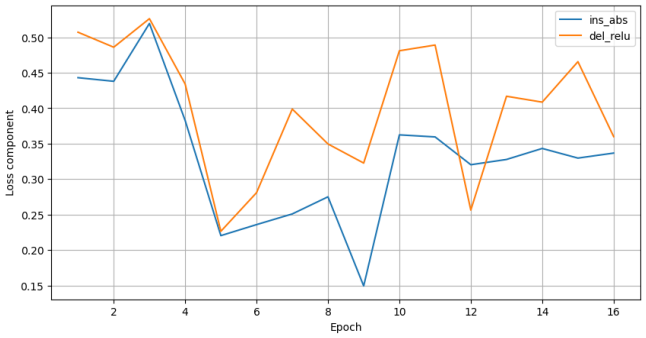

Fig. 9. Fidelity components during training.

To evaluate the effectiveness of the proposed explainability strategy, FusionMLP is compared with Grad-only, Attention-only, Avg, and Product fusion methods using insertion AUC, deletion AUC, and mask area metrics to assess explanation fidelity and compactness.

TABLE VIII. XAI FUSION COMPARISON

Method	Insertion AUC	Deletion AUC	Mask Area
Grad only	0.51	0.43	0.18
Attention only	0.56	0.39	0.21
Avg	0.63	0.34	0.15
Product	0.60	0.37	0.13
FusionMLP (Ours)	**0.7285**	**0.2681**	**0.12**

e) Evidence visualization and fusion (Grad-CAM, attention, fused output, mask): Fig. 10 shows, for each class, the input image, the Grad-CAM and proxy attention maps, and the fused mask. The fusion effectively combines local and global cues while remaining parsimonious, selecting approximately 12% of pixels, without min–max normalization in order to preserve probabilistic interpretation and avoid visual artifacts.

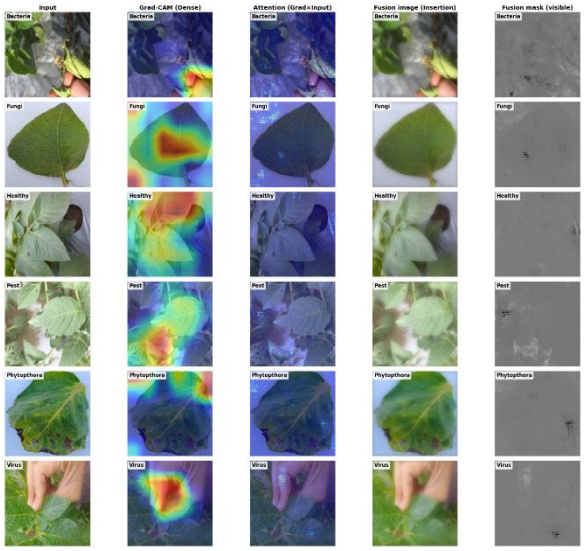

Fig. 10. Multi-source visual explanations showing the input image, Grad-CAM, Proxy Attention Map (Grad×Input), fused insertion map, and fusion mask.

For the Phytophthora class, the fused map accurately highlights lesion regions while suppressing irrelevant background areas. Compared with Grad-CAM alone, the fusion mechanism produces more compact and spatially coherent explanations by combining local lesion details with global contextual information.

D. Ablation studies

This section analyzes the impact of the various components and design choices of the proposed model through systematic ablation studies.

a) Part Classification : The ablation study shows that SwinV2-L outperforms DenseNet201 on the test set (98.72% vs. 95.96%), highlighting the effectiveness of attention-based contextual modeling. The proposed hybrid model achieves the best performance (98.87%, F1-score = 0.9887), confirming the complementarity between CNN local features and Transformer global representations under uncontrolled conditions.

b) Fusion module: The ablation study on fusion strategies demonstrates that the proposed FusionMLP consistently outperforms naive combinations, including Grad-CAM only (Grad), attention only (Attn), arithmetic mean (Avg), and element-wise product (Prod). In particular, FusionMLP achieves a substantially higher soft insertion AUC (≈ 0.73), along with a more balanced hard fidelity score (≈ 0.51), while strictly satisfying the area constraint (≈ 0.12). In contrast, baseline methods exhibit notably lower insertion performance and less favorable fidelity–area trade-offs, confirming the effectiveness of the learned soft-insertion-based fusion strategy.

c) Sensitivity analysis : To further analyze the robustness of the proposed fusion strategy, a sensitivity analysis was conducted on the main regularization parameters involved in the fusion objective function.

TABLE IX. SENSITIVITY ANALYSIS

Parameter	Tested values	Effect
AREA_TARGET	0.08 / 0.12 / 0.16	Fidelity–compactness trade-off
λ_2 (TV)	0.05 / 0.12 / 0.20	Spatial stabilization
λ_3 (Entropy)	0.005 / 0.01 / 0.02	Fidelity influence

The results in Table IX show that FusionMLP remains stable under moderate regularization variations while preserving compact explanatory masks.

E. Discussion: Robustness and Interpretability

The results show that a DenseNet201 + SwinV2 architecture achieves nearly 99% accuracy in real-world conditions while offering actionable explainability. Post-hoc fusion improves the stability and compactness of explanations, and despite room for improvement in soft fidelity, balanced hard fidelity and visual consistency confirm the value of the approach and open up prospects for improvement.

V. LIMITATIONS AND FUTURE WORK

Although DenseSwinLight demonstrates strong classification and explainability performance, the current evaluation remains limited to a unified potato leaf disease dataset and quantitative fidelity metrics. Future work will investigate cross-dataset validation, expert-centered evaluation, lesion localization, and deployment on edge devices.

VI. CONCLUSION

We presented DenseSwinLight, a hybrid CNN–Transformer approach for robust classification of potato leaf diseases in uncontrolled environments, coupled with a post-hoc explainability module using heatmap fusion. The model achieved 98.96% accuracy on test (TTA) with a macro F1 score of 0.9897. FusionMLP (with 65 parameters) produces parsimonious masks (area ≈ 0.12) and achieves an average fidelity of 0.498 (soft) and 0.513 (hard) according to

insertion/deletion metrics. These results suggest that constrained multi-source fusion is an effective strategy for balancing performance and interpretability, and provides a solid basis for future extensions in explainable AI.

REFERENCES

[1] B. Gülmez, "A Comprehensive Review of Convolutional Neural Networks based Disease Detection Strategies in Potato Agriculture," Potato Res., Aug. 2024, doi: 10.1007/s11540-024-09786-1.

[2] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady, and K. M. Hosny, "A hybrid Framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer," Complex Intell. Syst., vol. 11, no. 2, p. 142, Jan. 2025, doi: 10.1007/s40747-024-01764-x.

[3] H. Si, M. Li, W. Li, G. Zhang, M. Wang, F. Li, and Y. Li., "A Dual-Branch Model Integrating CNN and Swin Transformer for Efficient Apple Leaf Disease Classification," Agriculture, vol. 14, no. 1, Art. no. 1, Jan. 2024, doi: 10.3390/agriculture14010142.

[4] M. K. Gohil, A. Bhattacharjee, R. Rana, K. Lal, S. K. Biswas, N. Tiwari, and B. Bhattacharya, "A Hybrid Technique for Plant Disease Identification and Localisation in Real-time," arXiv preprint arXiv:2412.19682, Dec. 2024.

[5] F. Arshad, M. Mateen, S. Hayat, M. Wardah, Z. Al-Huda, Y. H. Gu, and M. A. Al-antari, "PLDPNet: End-to-end hybrid deep learning framework for potato leaf disease prediction," Alex. Eng. J., vol. 78, pp. 406–418, Sep. 2023, doi: 10.1016/j.aej.2023.07.076.

[6] H. Ghosh, I. S. Rahat, K. Shaik, S. Khasim, and M. Yesubabu, "Potato leaf disease recognition and prediction using convolutional neural networks," EAI Endorsed Trans. Scalable Inf. Syst., vol. 10, no. 6, Art. no. 6, Sep. 2023, doi: 10.4108/eetsis.3937.

[7] S. Mustofa, M. M. H. Munna, Y. R. Emon, G. Rabbany, and M. T. Ahad, "A Comprehensive Review on Plant Leaf Disease Detection Using Deep Learning," arXiv preprint arXiv:2308.14087, Aug. 2023.

[8] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, "Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT," arXiv preprint arXiv:2207.07919, Jul. 2022.

[9] S. Sundhar, R. Sharma, P. Maheshwari, S. R. Kumar, and T. S. Kumar, "Enhancing Leaf Disease Classification Using GAT-GCN Hybrid Model," arXiv preprint arXiv:2504.04764, Apr. 2025.

[10] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Comput. Electron. Agric., vol. 145, pp. 311–318, Feb. 2018, doi: 10.1016/j.compag.2018.01.009.

[11] A. Fuentes, S. Yoon, T. Kim, and D. S. Park, "Open Set Self and Across Domain Adaptation for Tomato Disease Recognition With Deep Learning Techniques," Front. Plant Sci., vol. 12, Dec. 2021, doi: 10.3389/fpls.2021.758027.

[12] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, Feb. 2021.

[13] Y. Wang, Y. Yin, Y. Li, T. Qu, Z. Guo, M. Peng, S. Jia, Q. Wang, W. Zhang, and F. Li, "Classification of Plant Leaf Disease Recognition Based on Self-Supervised Learning," Agronomy, vol. 14, no. 3, Art. no. 3, Mar. 2024, doi: 10.3390/agronomy14030500.

[14] S. Kar, K. Nagasubramanian, D. Elango, M. E. Carroll, C. A. Abel, A. Nair, D. S. Mueller, M. E. O'Neal, A. K. Singh, S. Sarkar, B. Ganapathysubramanian, and A. Singh, "Self-supervised learning improves classification of agriculturally important insect pests in plants," Plant Phenome J., vol. 6, no. 1, p. e20079, 2023, doi: 10.1002/ppj2.20079.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, Jun. 2021.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," J. Big Data, vol. 6, no. 1, p. 60, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[17] P. A. Dilip, K. Rameshbabu, K. P. Ashok, and S. A. Shivdas, "Bilinear interpolation image scaling processor for VLSI architecture," Int. J. Reconfigurable Embedded Syst., vol. 3, no. 3, pp. 104–113, Nov. 2014, doi: 10.11591/ijres.v3.i3.pp104-113.

[18] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 568–578, doi: 10.1109/ICCV48922.2021.00061.

[19] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin Transformer V2: Scaling Up Capacity and Resolution," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 11999–12009, doi: 10.1109/CVPR52688.2022.01170.

[20] D. P. Hughes and M. Salathe, "An Open Access Repository of Images on Plant Health to Enable the Development of Mobile Disease Diagnostics," arXiv preprint arXiv:1511.08060, Apr. 2016.

[21] N. H. Shabrina, S. Indarti, R. Maharani, D. A. Kristiyanti, and N. Prastomo, "A novel dataset of potato leaf disease in uncontrolled environment," Data Brief, vol. 52, p. 109955, 2024, doi: 10.1016/j.dib.2023.109955.

[22] D. Hirahara, E. Takaya, T. Takahara, and T. Ueda, "Effects of data count and image scaling on Deep Learning training," PeerJ Comput. Sci., vol. 6, p. e312, Nov. 2020, doi: 10.7717/peerj-cs.312.

[23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 30, 2017, arXiv: arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.

[24] Z. Zhang and M. R. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," Nov. 29, 2018, arXiv: arXiv:1805.07836. doi: 10.48550/arXiv.1805.07836.

[25] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," Jun. 20, 2023, arXiv: arXiv:2304.07288. doi: 10.48550/arXiv.2304.07288.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708. doi: 10.1109/CVPR.2017.243

[27] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[28] M. I. Rahman, "Fusion of Vision Transformer and Convolutional Neural Network for Explainable and Efficient Histopathological Image Classification in Cyber-Physical Healthcare Systems," Journal of Transformative Technologies and Sustainable Development, 2025, doi: 10.1007/s41314-025-00079-0.

[29] E. Hogea, D. M. Onchis, A. Coporan, A. M. Florea, and C. Istin, "ViTmiX: Vision Transformer Explainability Augmented by Mixed Visualization Methods," arXiv preprint arXiv:2412.14231, 2024.

[30] I. Lamaakal, C. Yahyati, Y. Maleh, K. El Makkaoui, and I. Ouahbi, "An explainable hybrid CNN–transformer model for sign language recognition on edge devices using adaptive fusion and knowledge distillation," Scientific Reports, vol. 16, 2026, doi: 10.1038/s41598-026-38478-8.

[31] M. Alshomrani, "An Explainable Hybrid CNN–Transformer Architecture for Malware Classification," Sensors, vol. 25, no. 15, p. 4581, 2025, doi: 10.3390/s25154581.

[32] D. C. Vitasoa, P. M. Randriamitsiry, H. Razafimahatratra, and T. Mahatody, "DenseSwinNet: A Hybrid CNN–Transformer Model for Robust Classification in Uncontrolled Environments," in Proc. IEEE ICSTCC, 2025, doi: 10.1109/ICSTCC66753.2025.11240329.