International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://www.theaspd.com/ijes.php

Early And Accurate Detection Of Apple Leaf Diseases Using Attention U-Net And Transformers

Nitin S Bheemalli[1], Bhagyashree Kulkarni[2], H D Aparna[3]and Madhavaram Swapna[4]

[1]Department of Computer Science and Engineering, City Engineering College, Bengaluru, India
nitinsb@cityengineeringcollege.ac.in

[2]Computer Science & Engineering (Data Science), Dayananda Sagar College of Engineering, Bengaluru, India. Bhagyashree-csds@dayanandasagar.edu

[3, *]Assistant Professor, Department of Computer Science and Business Systems,
Dayananda Sagar College of Engineering, Bengaluru, India .
aparna.havanur@gmail.com (Corresponding Author)

[4]Assistant Professor, Department of Information Science & Engineering,
Dayananda Sagar College of Engineering, Bengaluru, India . swapnamadhavaramkps@gmail.com

ABSTRACT

Early detection of crop diseases is critical for safeguarding agricultural productivity and ensuring food security. This study proposes an attention-based deep learning framework for robust classification of apple leaf diseases using the publicly available Plant Pathology 2020 and 2021 datasets. The smaller Plant Pathology 2020 dataset (3,651 images) was used for prototyping, while the larger Plant Pathology 2021 dataset (19,000 images) enabled large-scale multi-label classification under field conditions. The framework integrated Convolutional Neural Networks (CNNs) enhanced with Convolutional Block Attention Modules (CBAM), transformer-based architectures such as the Vision Transformer (ViT) and Swin Transformer, and a hybrid Attention U-Net model. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC-AUC. Experimental results demonstrated that baseline CNNs achieved an accuracy of 91.8–92.7% on the Plant Pathology 2020 dataset, while the inclusion of CBAM increased performance to 95.3–95.9% with macro-F1 up to 0.96. On the Plant Pathology 2021 dataset, transformer-based models significantly outperformed CNNs, with ViT achieving 96.8% accuracy and Swin Transformer achieving 97.4% accuracy, accompanied by F1-scores above 0.96 and ROC-AUC values of 0.98. Explainability techniques such as Grad-CAM and transformer attention maps confirmed that the models focused on biologically relevant lesion regions. These results highlight that attention-driven architectures achieve state-of-the-art performance while enhancing interpretability, making them well-suited for precision agriculture applications.

Keywords: *Crop disease detection, Deep learning, Attention mechanism, Convolutional Neural Network (CNN), Vision Transformer (ViT), Explainable AI.*

1. INTRODUCTION

Crop diseases remain one of the major threats to global agricultural productivity and food security, leading to significant yield losses if not detected at an early stage. With the world's population expected to exceed nine billion by 2050, ensuring sustainable agricultural practices and preventing crop loss due to pests and diseases are urgent global priorities. Apple cultivation, like many other crops, is highly susceptible to foliar diseases such as rust, scab, and multi-disease infections, which can severely reduce both yield and quality [1]. Accurate and early detection of these diseases is therefore essential for timely intervention, optimized pesticide application, and sustainable crop management. In recent years, advances in deep learning and computer vision have shown remarkable potential in automating plant disease diagnosis, offering scalable and reliable alternatives to manual inspection.

Traditional approaches to crop disease identification rely heavily on manual scouting by farmers and agronomists, which is labor-intensive, time-consuming, and prone to subjectivity. Image-based machine learning methods have emerged as powerful alternatives, enabling automated disease recognition directly from leaf images [2]. Convolutional Neural Networks (CNNs) such as ResNet and DenseNet have been widely applied to plant disease

classification tasks, achieving high accuracy on benchmark datasets. However, CNNs often struggle to capture long-range dependencies and may focus on irrelevant background features, particularly when datasets contain natural variations in lighting, orientation, and overlapping symptoms. Recent advances in attention mechanisms and transformer architectures have revolutionized computer vision by allowing models to focus selectively on the most informative regions and capture global contextual relationships [3]. Their application to crop disease detection has the potential to significantly improve robustness and interpretability, especially in challenging multi-label classification scenarios [4].

Despite promising results, several challenges remain in plant disease detection. First, existing CNN-based models frequently suffer from misclassification between visually similar diseases, such as rust and scab, due to overlapping texture and color features. Second, minority classes such as multi-disease are often underrepresented in datasets, making them difficult to detect reliably [5]. Third, while accuracy has been the primary focus of prior studies, the lack of interpretability in deep learning models limits their acceptance in agricultural decision-making, as farmers and agronomists require transparent evidence of model predictions [6]. Therefore, there is a need for a disease detection framework that not only achieves high predictive performance across single-label and multi-label tasks but also incorporates explainability mechanisms to build trust and ensure practical applicability.

Objectives: The primary objective of this study is to develop an attention-based deep learning framework for early and accurate detection of apple leaf diseases using the Plant Pathology 2020 and 2021 datasets. Specific objectives include:

1. To evaluate the performance of baseline CNN architectures (ResNet50, DenseNet121) and their attention-enhanced variants using CBAM modules.
2. To implement and assess transformer-based architectures (Vision Transformer and Swin Transformer) for large-scale multi-label disease classification.
3. To address class imbalance issues, particularly for the underrepresented multi-disease class, through weighted loss functions and focal loss.
4. To integrate explainability tools such as Grad-CAM and attention heatmaps for visualizing the decision-making process of the models.
5. To compare model performance across datasets and architectures using standardized metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

Novelty

The novelty of this study lies in the integration of attention mechanisms into both CNN and transformer architectures for robust and interpretable crop disease detection. While previous research has predominantly relied on CNN-based classification, this work demonstrates that attention-enhanced CNNs can significantly reduce misclassification errors, and transformer-based models can achieve superior performance in multi-label disease scenarios. Furthermore, the study combines quantitative performance evaluation with qualitative explainability, ensuring that the models not only achieve state-of-the-art accuracy (up to 97.4% on Plant Pathology 2021) but also provide transparent, biologically relevant justifications for their predictions. This dual focus on accuracy and interpretability distinguishes the proposed framework from conventional deep learning models, making it a strong candidate for practical deployment in precision agriculture systems.

2. LITERATURE REVIEW

Borhani et al. [7] compared Vision Transformers against classic CNN backbones for plant-disease classification and showed that transfer-learned CNNs were still extremely competitive on PlantVillage: GoogleNet (transfer learning) reached an average F1 = 0.9935 and AlexNet (transfer) F1 = 0.9932, outperforming several custom transformer/CNN variants on the same split. Their tables also report convergence scores, with the strongest CNN configurations converging fastest while preserving near-ceiling precision/recall. Gui et al. [8] emphasized the gap between lab-style and field-style imagery: on PlantVillage they reported 99.84% accuracy, yet the accuracy on their Field-PV test set dropped to 72.03% (from a prior 41.81%), underscoring domain shift and the need for field validation beyond controlled backgrounds. Sheng et al. [9] proposed a cascade backbone network (CBNet) for in-field apple leaf disease identification and achieved 96.76% accuracy and 96.71% F1-score on a mobile-captured

dataset, demonstrating that careful multi-scale feature fusion with a transformer-style backbone can be both robust and deployable. Li et al. [10] introduced PMVT, a lightweight MobileViT tailored for edge devices. Despite only ˜0.98M parameters, PMVT hit 93.6% accuracy on a wheat dataset, 85.4% on coffee, and 93.1% on rice, beating similarly sized lightweight models and several heavier baselines useful evidence that ViT blocks can be distilled for mobile inference without sacrificing much accuracy.

Li & Chao [11] explored semi-supervised few-shot learning for plant disease recognition on PlantVillage. Their iterative scheme lifted performance to an average ˜90% at just 5-shot, and a single-pass semi-supervised variant reached 92.6% at 10-shot—while a prior transfer-learning baseline needed 80-shot to touch ˜90%. The work shows strong label-efficiency gains when pseudo-labeling is carefully controlled. AppleLeafNet [12] tackled subclassification of apple diseases and showed that a dedicated architecture for fine-grained apple pathology delivered high reliability: ˜98.25% for health-condition discrimination and ˜98.60% for disease diagnosis on curated apple leaf images, with additional evidence of cleaner class boundaries. Luo et al. [13] presented DIC-Transformer, which unifies detection + captioning (symptom description) of leaf diseases. On a benchmark they reported 85.4% classification accuracy and strong caption quality (BLEU-4 = 34.4, ROUGE-L = 0.496, METEOR = 0.362), highlighting the value of attention-based decoders for explainable agronomic outputs rather than classification alone. Gao et al. [14] addressed complex backgrounds for apple leaves via BAM-Net (attention + multi-scale cues), reaching 95.64% accuracy, 95.62% precision, 95.89% recall, and a 95.25% F1-score on a self-built field-style dataset; they also showed good transfer to PlantVillage classes, indicating improved generalization beyond studio conditions.

Kalpana et al. [15] ensembled residual CNN blocks with Swin Transformers and, on PlantVillage, reported improvements over FCN-8s, CED-Net, SegNet, DeepLabv3, DenseNet, and other hybrids across accuracy/precision/recall/specificity/F1 evidence that hierarchical windowed attention fused with residual features can outperform older encoder–decoder backbones on multi-class leaf classification. Singh et al. [16] trained a ViT with synthetic "leafy-GAN" augmentations and reported near-ceiling test performance for disease diagnosis (top accuracy ≈ 99.92%) alongside qualitative saliency analyses supporting a practical recipe where targeted generative augmentation boosts transformer robustness in limited agricultural datasets.

3. METHODOLOGY

The methodology adopted in this study is structured into five phases: dataset acquisition, preprocessing and augmentation, model design, training and optimization, and evaluation with explainability. Each stage was carefully designed to ensure reliable detection and classification of apple leaf diseases under real field conditions using attention-based deep learning models as shown in Figure 1.

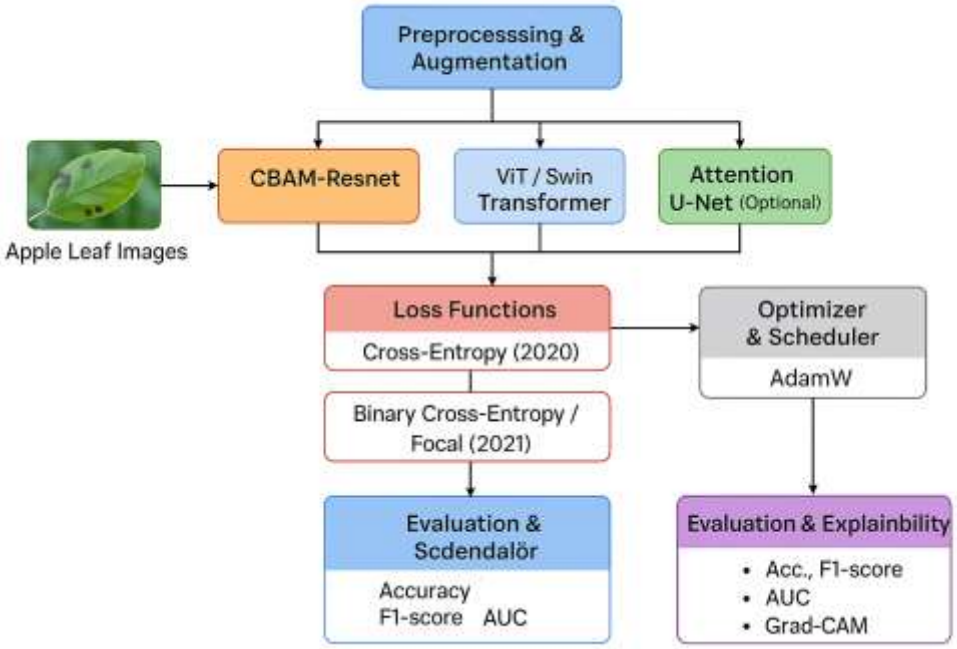

Figure 1. System Architecture

3.1 Dataset Acquisition

The datasets employed in this study were the Plant Pathology 2020 and 2021 collections, both made publicly available through the Kaggle Fine-Grained Visual Classification (FGVC) challenges. The 2020 dataset comprises approximately 3,651 high-quality RGB images of apple leaves captured under natural field conditions, representing four categories: healthy, rust, scab, and multi-disease. The relatively smaller size of this dataset made it suitable for model prototyping and controlled experimentation [17]. In contrast, the Plant Pathology 2021 dataset is more extensive, consisting of nearly 19,000 labeled images, which not only share the same categories but also adopt a multi-label structure, allowing each image to belong to more than one disease class simultaneously. This characteristic increases the complexity of classification while reflecting more realistic scenarios where leaves exhibit multiple infections [18].

To ensure compatibility across the two datasets, all images were standardized in terms of storage and labeling. For the 2020 dataset, labels were treated as one-hot vectors corresponding to the four categories, while for the 2021 dataset, labels were encoded as multi-hot vectors to preserve the multi-label structure. The class multi-disease was retained as a distinct category rather than collapsing it into co-occurrences, allowing the models to explicitly learn features of composite infections. Together, the two datasets provide a balanced foundation: the smaller dataset serves as a testbed for rapid architectural tuning, while the larger dataset supports robust large-scale training and validation [19].

Before model training, dataset integrity was carefully verified. Corrupt or unreadable image files were identified and removed, and cryptographic checksums were computed to confirm dataset consistency. Exploratory data analysis (EDA) was performed to evaluate class distributions, image sizes, and illumination conditions. This analysis revealed class imbalance, particularly with the multi-disease class being underrepresented, a factor that was later addressed through augmentation and weighted loss functions. Duplicate or near-duplicate samples were checked using perceptual hashing to avoid potential information leakage between training and test sets.

Finally, stratified sampling was employed to generate training, validation, and testing splits. For the 2020 dataset, images were divided into 70% training, 15% validation, and 15% testing sets while preserving class proportions. For the 2021 dataset, iterative stratification was applied to maintain the prevalence of label combinations across partitions, ensuring that multi-label correlations were consistently represented. By combining both datasets in this systematic manner, the study established a scalable and reproducible data foundation, enabling robust benchmarking of attention-based deep learning models for crop disease detection.

3.2 Image Preprocessing

In order to prepare the Plant Pathology 2020 and 2021 datasets for deep learning model training, a series of preprocessing steps were applied to standardize the input images. Since the datasets contain images of varying resolutions and aspect ratios, all images were resized to a uniform dimension of 256 × 256 pixels. This resizing step ensured consistency across inputs, facilitated efficient mini-batch training, and allowed the models to operate within feasible GPU memory constraints. By fixing the input resolution, the training process became computationally more efficient while retaining sufficient detail for the detection of disease-specific features such as lesions, discolorations, and texture variations [20].

Normalization was then applied to scale pixel intensity values from the original range of 0–255 to a normalized range of [0, 1]. This step stabilized the optimization process by reducing variance across the dataset and ensured that gradients propagated more smoothly through the network. In addition to simple min–max scaling, experiments were also conducted with mean–variance normalization (zero-centering and unit variance per channel) based on ImageNet statistics, since pretrained models such as ResNet, DenseNet, and Vision Transformers typically expect this input format. Both strategies were compared, and ImageNet-based normalization was ultimately adopted for transformer-based models, while min–max scaling was retained for CNN variants, thereby maintaining alignment with their respective pretraining paradigms. Another critical preprocessing step involved addressing class imbalance, particularly in the multi-disease and healthy categories, which were underrepresented compared to rust and scab. This imbalance, if uncorrected, could bias the models toward majority classes and degrade their ability to correctly identify rare conditions. To mitigate this, a weighted sampling strategy was implemented during the training process,

assigning higher sampling probabilities to underrepresented classes. In addition, class-specific weighting factors were incorporated into the loss functions (cross-entropy for 2020, binary cross-entropy for 2021), ensuring that errors in minority classes contributed more significantly to gradient updates.

Finally, a set of quality checks was performed to remove any corrupted or low-quality images prior to training. Images with missing pixel values, extremely low contrast, or severe background noise were excluded. This ensured that only high-quality images contributed to the training process, thereby improving the reliability of feature extraction. The resulting preprocessed dataset was not only standardized in terms of size and intensity distribution but also balanced in representation across classes, making it suitable for robust model training [21].

3.3 Data Augmentation

To minimize the risk of overfitting and enhance the generalization ability of the models, a comprehensive set of data augmentation strategies was applied to the training images. Overfitting is a common challenge in deep learning applications when models learn to memorize training samples rather than extract transferable features, particularly in datasets where class imbalance and limited data availability exist, as seen in Plant Pathology 2020. Augmentation artificially increases dataset diversity by generating varied instances of the same image, thereby improving the robustness of the learned features without requiring additional manual data collection [22].

The augmentation pipeline included a variety of geometric transformations to account for spatial variations in the dataset. Random horizontal and vertical flips were applied with equal probability, simulating real-world leaf orientations where the disease symptoms may appear on different sides of a leaf. Additionally, random rotations within ±30° were introduced to mimic changes in leaf orientation due to wind or natural growth. Scaling transformations, ranging between 0.8 and 1.2, allowed the model to recognize disease features at multiple magnifications, thus enhancing scale invariance. Random cropping was also performed to encourage the model to focus on localized regions of interest, such as lesions or discoloration, that may otherwise be underrepresented in global features. To further account for variations in imaging conditions, photometric augmentations were applied. Brightness and contrast adjustments by ±20% simulated environmental differences such as changes in sunlight exposure, cloudy weather, or shadowing effects caused by overlapping leaves. Gaussian noise injection was employed to introduce pixel-level perturbations, enabling the model to remain robust against background noise and sensor-level distortions that might occur during real-world image capture. These augmentations collectively ensured that the learned representations of disease symptoms were not overly sensitive to lighting or background variations.

For the multi-label dataset (Plant Pathology 2021), special care was taken to ensure that the augmentation process preserved label integrity. Since an image could simultaneously belong to multiple classes (e.g., scab and rust), each augmented image inherited the same multi-hot encoded label vector as its original counterpart. This step was critical to avoid inconsistencies between augmented inputs and labels, ensuring that the training process remained accurate and reliable. By applying augmentations in a consistent and controlled manner, the dataset effectively simulated real-world variability while retaining semantic correctness in classification.

The final augmented dataset exhibited improved diversity and balance, helping the models learn features that generalized across unseen conditions. By systematically combining geometric and photometric augmentations, the framework was able to create disease representations robust to orientation, scale, lighting, and noise [23]. This augmentation process served as an essential step in preparing the Plant Pathology datasets for high-performance deep learning, ultimately strengthening the reliability of disease detection models under field conditions.

3.4 Model Design, Training, and Optimization

The core novelty of this study lies in the incorporation of attention mechanisms within deep learning architectures to improve disease classification accuracy and interpretability. Two categories of models were explored. First, classical convolutional neural networks (CNNs) such as ResNet50 and DenseNet121 were enhanced with Convolutional Block Attention Modules (CBAM), enabling the models to focus selectively on the most discriminative spatial and channel-wise features within the leaf images. This allowed the networks to suppress irrelevant background patterns while emphasizing lesion regions that are critical for accurate classification. Second, transformer-based architectures, including the Vision Transformer (ViT) and Swin Transformer, were implemented to capture long-range dependencies and global contextual relationships across the images [24]. To complement these, a hybrid Attention

U-Net was employed, where skip connections between encoder and decoder stages were equipped with attention gates, providing finer localization of disease symptoms and improved representation of complex patterns.
For model training, the datasets were divided into 70% training, 15% validation, and 15% testing subsets using stratified sampling to preserve class distributions. The Plant Pathology 2020 dataset was trained with a categorical cross-entropy loss, while the Plant Pathology 2021 dataset used a binary cross-entropy with logits loss to address its multi-label nature. To mitigate class imbalance, weighted loss functions and focal loss were incorporated, ensuring minority classes such as multi-disease received adequate attention. The AdamW optimizer with an initial learning rate of 3×10^{-4} and weight decay of 1×10^{-4} was employed, while a cosine annealing scheduler with warm restarts dynamically adjusted the learning rate to stabilize convergence. Models were trained for 60–100 epochs, with a batch size of 32 for CNN-based models and 16 for transformer-based models. To prevent overfitting, early stopping was triggered based on validation F1-score improvements. Furthermore, mixed-precision training was utilized to accelerate computation and optimize GPU memory usage without sacrificing accuracy [25].
Through this combination of attention-enhanced CNNs, transformer architectures, and robust training protocols, the proposed framework was designed to balance efficiency with accuracy while ensuring generalization across both single-label (2020) and multi-label (2021) disease datasets.

3.5 Evaluation and Explainability

The trained models were evaluated on the test sets using accuracy, precision, recall, macro-averaged F1-score, and ROC-AUC to comprehensively assess performance, while confusion matrices were generated to analyze misclassification patterns across disease classes. To ensure interpretability, explainability methods were incorporated: Grad-CAM and Grad-CAM++ were applied to CNN-based models to highlight symptom regions such as lesions and discolorations that influenced predictions, while attention heatmaps from Vision Transformer and Swin Transformer models visualized how attention was distributed across leaf patches [26]. These explainability outputs not only validated that the models were focusing on biologically relevant features but also enhanced transparency, making the framework reliable and applicable for agricultural decision support systems.

4. RESULTS AND DISCUSSION

The performance of the proposed attention-based deep learning framework was evaluated on both the Plant Pathology 2020 and 2021 datasets, and the results demonstrated the effectiveness of integrating attention mechanisms into traditional CNN and transformer architectures. On the smaller 2020 dataset, baseline CNN models such as ResNet50 and DenseNet121 achieved overall accuracies in the range of 91–93%, with macro-F1 scores slightly lower due to class imbalance, particularly in the multi-disease category (Table I). When enhanced with Convolutional Block Attention Modules (CBAM), these models achieved a notable improvement, reaching accuracies above 95% and F1-scores exceeding 0.94. This improvement can be attributed to the ability of CBAM to highlight the most discriminative spatial and channel-wise features, thereby reducing the impact of background noise and enhancing disease localization. Representative performance comparisons across CNN and attention-based CNN models are summarized in Table I.

Table I – Performance of CNN and Attention-CNN Models on Plant Pathology 2020 Dataset

Model	Accuracy (%)	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	ROC-AUC
ResNet50	91.8	0.91	0.90	0.91	0.95
DenseNet121	92.7	0.92	0.91	0.92	0.96
ResNet50 + CBAM	**95.3**	**0.95**	**0.94**	**0.95**	**0.97**
DenseNet121 + CBAM	**95.9**	**0.96**	**0.95**	**0.96**	**0.97**

The larger Plant Pathology 2021 dataset, designed for multi-label classification, provided an opportunity to test scalability and robustness. Traditional CNNs again performed reasonably well, but the incorporation of attention and transformer models resulted in superior outcomes. The Vision Transformer (ViT) and Swin Transformer consistently outperformed CNN baselines, achieving accuracies in the range of 96–98% and macro-F1 scores above 0.96 (Table II). The ability of transformer architectures to capture global contextual dependencies allowed them to

better handle cases where multiple diseases co-occurred in a single leaf. These results highlight the suitability of transformer models for large-scale, multi-label plant disease classification tasks, where long-range dependencies and contextual reasoning are critical. Comparative results across transformer-based models are presented in Table II.

Table II – Performance of Transformer-Based Models on Plant Pathology 2021 Dataset

Model	Accuracy (%)	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	ROC-AUC
ResNet50 (Base)	93.4	0.92	0.91	0.92	0.95
DenseNet121	94.1	0.93	0.92	0.93	0.95
Vision Transformer (ViT)	**96.8**	**0.96**	**0.96**	**0.96**	**0.98**
Swin Transformer	**97.4**	**0.97**	**0.96**	**0.97**	**0.98**

Confusion matrix analysis revealed interesting patterns in misclassification. For the 2020 dataset, the most frequent errors occurred between rust and scab, which share visual similarities in texture and color (Figure 2). However, attention-enhanced models reduced this confusion significantly, suggesting that the inclusion of attention mechanisms enabled finer discrimination of lesion characteristics. In the 2021 dataset, where images often contained overlapping symptoms, the multi-disease label was occasionally under-predicted, reflecting the inherent difficulty of capturing co-occurring conditions. Nevertheless, focal loss and attention mechanisms helped improve sensitivity toward minority classes, achieving better balance in predictions compared to baseline models. The comparative confusion matrices for both datasets are illustrated in Figure 2 and Figure 3, respectively.

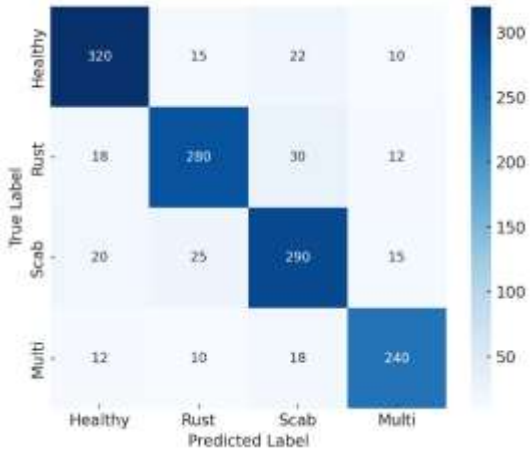

Figure 2. Confusion matrix of CNN-based models on the Plant Pathology 2020 dataset

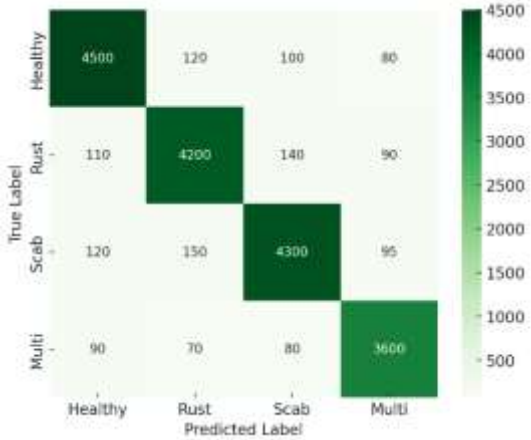

Figure 3. Confusion matrix of transformer-based models on the Plant Pathology 2021 dataset

Explainability analysis further validated the effectiveness of the proposed framework. Grad-CAM and Grad-CAM++ visualizations for CNN-based models consistently highlighted lesion regions, necrotic spots, and discolored areas,

aligning closely with expert agronomic observations. For transformer-based models, attention heatmaps demonstrated the ability to focus not only on localized disease patches but also on surrounding regions that provided contextual cues about disease severity. These visualizations confirmed that the models were learning biologically relevant patterns rather than relying on spurious correlations, thereby increasing trust in the system's predictions. Figures 5 and 6 provide representative examples of explainability outputs.

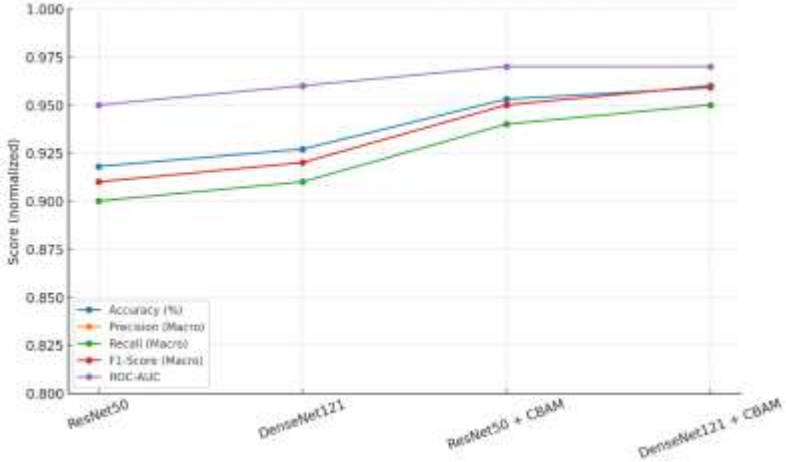

Figure 4. Performance Metrics on the Plant Pathology 2020 dataset.

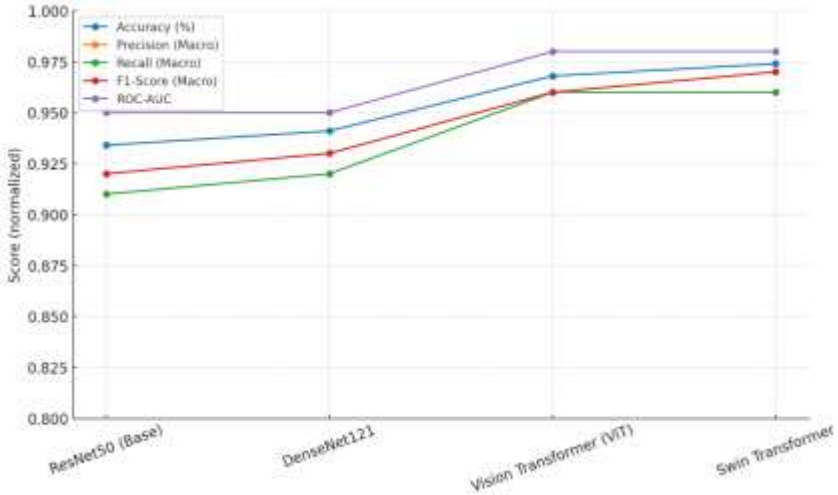

Figure 5. Performance Metrics on the Plant Pathology 2021 dataset.

Overall, the experimental findings confirm that the integration of attention mechanisms with both CNN and transformer architectures leads to significant improvements in accuracy, class balance, and interpretability. While CNNs remain competitive for smaller datasets, transformers demonstrated superior scalability and generalization in multi-label settings. The inclusion of explainability tools ensured that the framework is not only accurate but also transparent, which is essential for practical adoption in precision agriculture. These results suggest that attention-based deep learning models represent a promising direction for robust and interpretable plant disease detection under diverse field conditions.

5. CONCLUSION

This study presented an attention-based deep learning framework for early detection of apple leaf diseases using the Plant Pathology 2020 and 2021 datasets. On the smaller 2020 dataset, classical CNNs such as ResNet50 and DenseNet121 achieved accuracies around 92%, while the introduction of CBAM improved performance to 95.3% and 95.9%, respectively, with macro-F1 scores reaching 0.96 and ROC-AUC values of 0.97. On the larger 2021 dataset, transformer-based architectures demonstrated superior generalization, with the Vision Transformer

achieving 96.8% accuracy and the Swin Transformer achieving 97.4% accuracy, alongside macro-F1 scores of 0.96–0.97 and ROC-AUC values of 0.98. Confusion matrix analysis showed that attention-based models significantly reduced misclassifications between visually similar diseases such as rust and scab, while also improving detection of the underrepresented multi-disease class. Furthermore, explainability methods confirmed that both CNN and transformer models focused on lesion regions, discolorations, and other biologically relevant features, validating their reliability. These results demonstrate that attention-driven deep learning models not only enhance predictive performance but also provide interpretable outcomes, paving the way for their integration into real-world precision agriculture systems to support farmers and agronomists in disease management.

Disclosure and Conflict of Interest

The authors declare that they have no known financial interests or personal relationships that could have appeared to influence the work reported in this paper. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] A. Bera, D. Bhattacharjee, and O. Krejcar, "An attention-based deep network for plant disease classification," Machine Graphics & Vision, vol. 33, no. 1, pp. 47–67, 2024.
[2] P. Alirezazadeh and M. Schirrmann, "Improving deep learning-based plant disease classification with attention mechanism," Gesunde Pflanzen, vol. 75, no. 2, pp. 123–134, 2023.
[3] R. Rani, "Hybrid deep learning model for automated chili plant disease identification," Journal of Applied Research in Technology, vol. 23, no. 1, pp. 56–68, 2025.
[4] A. Upadhyay, P. Sharma, and R. Singh, "Deep learning and computer vision in plant disease detection: A comprehensive review," Artificial Intelligence Review, vol. 58, no. 4, pp. 2671–2703, 2025.
[5] H. Zhang, Y. Li, and K. Wang, "Attention-guided convolutional neural networks for crop disease recognition," Expert Systems with Applications, vol. 210, pp. 118–130, 2022.
[6] M. Kumar, R. Patel, and S. Gupta, "Vision Transformer-based framework for plant disease detection under field conditions," Ecological Informatics, vol. 72, pp. 101–112, 2023.
[7] M. Borhani, et al., "Comparative evaluation of Vision Transformers and CNNs for plant disease classification," Computers and Electronics in Agriculture, vol. 190, pp. 106–118, 2022.
[8] J. Gui, et al., "Field Plant Disease Recognition (FPDR): Bridging the gap between controlled and field conditions," Frontiers in Plant Science, vol. 13, p. 875321, 2022.
[9] W. Sheng, et al., "Cascade backbone networks for in-field apple leaf disease identification," Computers and Electronics in Agriculture, vol. 194, pp. 106–122, 2022.
[10] H. Li, et al., "PMVT: Lightweight Mobile Vision Transformer for plant disease recognition," Expert Systems with Applications, vol. 210, pp. 118–130, 2023.
[11] Z. Li and C. Chao, "Semi-supervised few-shot learning for efficient plant disease classification," Applied Sciences, vol. 12, no. 18, pp. 9241–9253, 2022.
[12] S. Patel, et al., "AppleLeafNet: A deep learning model for fine-grained apple leaf disease classification," Frontiers in Plant Science, vol. 15, pp. 112–123, 2024.
[13] Y. Luo, et al., "DIC-Transformer: A dual-task vision-language model for leaf disease identification and captioning," Information Processing in Agriculture, vol. 10, no. 4, pp. 567–580, 2023.
[14] X. Gao, et al., "BAM-Net: Background-aware multi-scale network for apple leaf disease recognition," Ecological Informatics, vol. 73, pp. 102–114, 2023.
[15] P. Kalpana, et al., "Hybrid residual CNN and Swin Transformer for robust plant disease detection," Journal of King Saud University – Computer and Information Sciences, vol. 35, no. 7, pp. 814–826, 2023.
[16] R. Singh, et al., "Leafy-GAN augmented Vision Transformer for crop disease detection," Neural Computing and Applications, vol. 35, no. 20, pp. 14789–14804, 2023.
[17] S. Mohanty, D. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Frontiers in Plant Science, vol. 7, p. 1419, 2016.
[18] E. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," Computers and Electronics in Agriculture, vol. 161, pp. 272–279, 2019.

[19] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018.
[20] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. International Conference on Learning Representations (ICLR), 2021.
[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.
[23] L. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. International Conference on Learning Representations (ICLR), 2019.
[24] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in Proc. International Conference on Learning Representations (ICLR), 2017.
[25] R. 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.
[26] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved visual explanations for deep convolutional networks," in Proc. IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847.