



Title Page

LOC Page

This book is dedicated to all the researchers, professionals, and students who are continuously trying their best for the advancement of technologies for the betterment of humanity.



Contents

CHAPTER 1	■ Precision Agriculture for Jujube: Swin Transformer Based Early Identification of Insect Pest Damage Using Explainable AI	1
	Md FARHAN SHAHRIYAR and ABDULLAH Md RAIHAN CHY	
1.1	INTRODUCTION	2
1.2	LITERATURE REVIEW	3
1.3	METHODOLOGY	4
1.3.1	Dataset Description	5
1.3.2	Data Preprocessing and Augmentation	5
1.3.3	Swin Transformer Architecture	5
1.3.4	Loss Function	6
1.3.5	Training Configuration	6
1.3.6	Evaluation Metrics	6
1.3.7	Explainability Framework	7
1.4	RESULTS AND DISCUSSION	7
1.4.1	Model Performance Comparison	7
1.4.2	Per-Class Performance Analysis	7
1.4.3	Confusion Matrix Analysis	8
1.4.4	ROC and Precision-Recall Analysis	9
1.4.5	Training Dynamics	9
1.4.6	Explainability Analysis	10
1.4.7	Architectural Insights	11
1.4.8	Practical Implications	11
1.5	CONCLUSION	12



Precision Agriculture for Jujube: Swin Transformer Based Early Identification of Insect Pest Damage Using Explainable AI

Md Farhan Shahriyar

Department of Computer Science and Engineering, Hajee Mohammad Danesh Science & Technology University, Basherhat. Research & Development, Zogbiyog. Email: farhanshahriyar.cse1@gmail.com.

Abdullah Md Raihan Chy

Department of Computer Science and Engineering, Southern University Bangladesh, Chattogram, Bangladesh. Research & Development, Zogbiyog. Email: raihanchowdhuryengineer@gmail.com.

CONTENTS

1.1	Introduction	2
1.2	Literature Review	3
1.3	Methodology	4
1.3.1	Dataset Description	5
1.3.2	Data Preprocessing and Augmentation	5
1.3.3	Swin Transformer Architecture	5
1.3.4	Loss Function	6
1.3.5	Training Configuration	6
1.3.6	Evaluation Metrics	6
1.3.7	Explainability Framework	7
1.4	Results and Discussion	7
1.4.1	Model Performance Comparison	7
1.4.2	Per-Class Performance Analysis	7
1.4.3	Confusion Matrix Analysis	8

1.4.4	ROC and Precision-Recall Analysis	9
1.4.5	Training Dynamics	9
1.4.6	Explainability Analysis	10
1.4.7	Architectural Insights	11
1.4.8	Practical Implications	11
1.5	Conclusion	12

Early detection of insect pest damage in jujube cultivation is critical for timely pest management. This chapter proposes a Swin Transformer based framework for classifying jujube leaf conditions into three categories: Healthy, Miner-damaged, and Weevil-damaged. We compare four vision transformer architectures: Swin Transformer (Base, Small, Tiny) and Vision Transformer (ViT-Base). Swin Transformer Small achieved the highest test accuracy of 98.90% with F1-scores of 100%, 98.3% and 98.4% for the Healthy, Miner and Weevil classes, and AUC-ROC of 1.000 across all classes. The framework offers an accurate and interpretable solution for precision pest management in jujube orchards.

Keywords: Jujube; Swin Transformer; Pest Detection; Explainable AI; Precision Agriculture

1.1 INTRODUCTION

Jujube (*Ziziphus jujuba*), commonly known as Chinese date, represents a commercially significant fruit crop cultivated extensively across Asia, particularly in China, India, and the Middle East [1]. The fruit possesses substantial nutritional and medicinal value, making its protection from pest damage economically imperative. Among the various insect pests affecting jujube cultivation, leaf miners and weevils constitute major threats that cause significant yield reductions and quality deterioration [2]. Traditional pest monitoring approaches relying on manual visual inspection by agronomists are labor-intensive, time-consuming, and often inconsistent, particularly in large-scale orchards [3].

The integration of artificial intelligence (AI) in agricultural practices has emerged as a transformative solution for automated crop health monitoring. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable success in plant disease and pest classification tasks [4, 5, 6]. However, recent advancements in Vision Transformers have shown superior performance in capturing long-range dependencies within images through self-attention mechanisms [7].

The Swin Transformer architecture, introduced by Liu et al. [8], represents a significant advancement in vision transformers by employing shifted window-based self-attention that enables linear computational complexity while maintaining hierarchical feature representation. This architecture addresses the quadratic complexity limitation of standard Vision Transformers while preserving the ability to model global contextual information. The shifted window mechanism partitions features into fixed windows with cyclic shifts, enabling efficient local self-attention and cross-window information exchange for hierarchical feature integration.

Explainable AI (XAI) has gained prominence in agricultural applications by addressing

the “black box” nature of deep learning models [9]. Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) provide visual explanations by highlighting image regions that influence model predictions, thereby fostering trust and enabling validation of model decisions [10]. The integration of XAI ensures that automated pest detection systems remain interpretable and trustworthy for practical deployment.

This chapter presents a comprehensive evaluation of vision transformer architectures for jujube leaf pest damage classification. The primary contributions include:

1. Development of a Swin Transformer-based classification framework for three-class pest damage identification,
2. Comparative analysis of Swin Transformer variants (Base, Small, Tiny) against standard Vision Transformer,
3. Integration of explainable AI techniques for model interpretability and demonstration of the proposed system’s efficacy for precision agriculture applications.

The remainder of this chapter is organized as follows. **Section 1.2** discusses related works in the area of pest detection. The proposed deep learning-based jujube pest identification model and dataset are given in **Section 1.3**, followed by discussions on the classification performances of the proposed model in **Section 1.4**. Finally, **Section 1.5** concludes the chapter.

1.2 LITERATURE REVIEW

Deep learning frameworks have increasingly been adopted for pest classification in agricultural settings [11]. Methods leveraging mobile vision-based approaches have demonstrated promising results for field deployment. Kumar et al. [3] proposed AI-SCAN, combining residual convolutional networks with Swin Transformer architecture for plant leaf disease diagnosis, achieving enhanced reliability and accuracy through hybrid mechanisms. The system utilized both spatial and self-attention methods to recognize diseases accurately.

HyPest-Net, a hybrid architecture integrating CNNs with channel attention mechanisms and ViT-B/16 modules, achieved 95% accuracy on rice pest datasets by resolving challenges posed by visually similar species and background clutter [7]. These studies demonstrate the potential of transformer-based approaches for agricultural pest classification.

Vision Transformers have demonstrated superior capability in modeling long-range dependencies essential for complex pattern recognition in agricultural images [12]. Unlike CNNs that rely on local receptive fields, ViTs employ self-attention mechanisms to learn relationships between all image patches simultaneously. However, the quadratic computational complexity of standard ViTs limits their scalability for high-resolution agricultural imagery [8].

The Swin Transformer addresses these limitations through hierarchical feature representation and window-based self-attention. Yu et al. [13] proposed ST-CFI combining Swin Transformer with convolutional feature integration, representing substantial advancement in early and accurate plant disease detection. Liu et al. [14] developed an Efficient Swin Transformer demonstrating excellent overall performance in plant disease recognition tasks.

Research specifically targeting jujube pest detection remains limited. Wu et al. [15] explored hyperspectral remote sensing combined with clustering algorithms for spider mite

recognition in jujube trees, achieving 96.13% overall accuracy. Erdinç et al. [16] applied YOLOv5 variants for detecting *Metcalfa pruinosa* adults on jujube plants through field images. However, these studies focused on specific pest types using conventional approaches, leaving transformer-based methods unexplored for comprehensive jujube pest classification.

The deployment of AI in agricultural decision-making necessitates interpretable models that practitioners can trust [9]. XAI techniques clarify which features influence model predictions, enabling validation against domain knowledge [17]. Grad-CAM has emerged as a vital tool for providing visual explanations of Vision Transformer predictions [10].

Transfer learning from ImageNet-pretrained models has proven effective for agricultural applications where labeled data is limited [18]. Al Sahili et al. demonstrated that domain-specific pretrained models achieve state-of-art performance, with AgriNet models trained on 160,000 agricultural images achieving 88–94% accuracy across architectures [18].

1.3 METHODOLOGY

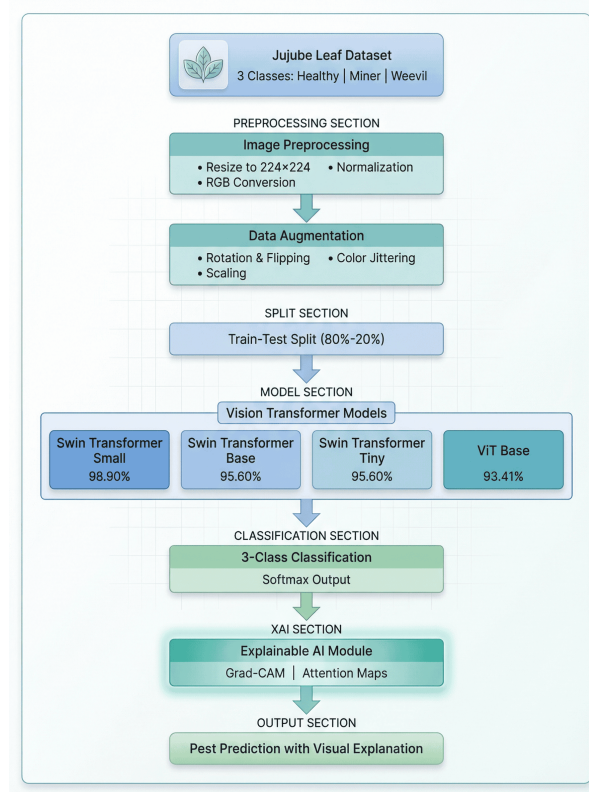


Figure 1.1: Proposed framework for jujube leaf pest detection.

The overall architecture of the proposed pest detection system is depicted in Fig. 1.1. Raw jujube leaf images belonging to three categories (Healthy, Miner, Weevil) undergo pre-processing operations including dimension standardization to 224×224 pixels, color space conversion to RGB, and statistical normalization. To improve model robustness, various augmentation strategies such as geometric transformations and photometric distortions are

applied during training. Four transformer-based architectures are subsequently trained and evaluated. Classification outputs pass through a softmax layer to produce probability distributions across pest categories. The framework incorporates an interpretability module utilizing Grad-CAM to generate visual explanations accompanying each prediction.

1.3.1 Dataset Description

This chapter utilizes a curated collection of jujube leaf photographs [19] organized into three distinct categories. The Healthy class contains images of undamaged foliage exhibiting normal coloration and structure. The Miner class comprises leaves showing characteristic damage patterns from leaf-mining insect larvae. The Weevil class includes samples displaying feeding damage typical of weevil infestations. Image acquisition occurred across multiple orchard locations under diverse illumination conditions to ensure dataset variability and model generalizability.

1.3.2 Data Preprocessing and Augmentation

Input images undergo spatial rescaling to dimensions of 224×224 pixels, matching the expected input resolution of pretrained transformer networks. The training subset receives additional transformations to artificially expand dataset diversity and mitigate overfitting:

- Horizontal and vertical mirror operations with 50% probability
- Rotational perturbations within a $\pm 20^\circ$ range
- Photometric adjustments modifying brightness, contrast, and saturation by factors up to 0.2, with hue shifts up to 0.1
- Spatial translations up to 10% of image dimensions

Pixel intensity values are standardized using channel-wise statistics derived from the ImageNet corpus (mean values: 0.485, 0.456, 0.406; standard deviation values: 0.229, 0.224, 0.225). This normalization aligns input distributions with those encountered during original model pretraining. The complete dataset undergoes stratified partitioning into training (70%), validation (15%), and testing (15%) subsets, preserving class proportions across all splits.

1.3.3 Swin Transformer Architecture

The Swin Transformer architecture constructs hierarchical feature representations through a novel windowed self-attention mechanism [8]. Rather than computing attention globally across all spatial positions, this approach restricts attention computations to local windows, substantially reducing computational requirements. Consecutive transformer blocks alternate between standard window partitioning (W-MSA) and shifted window configurations (SW-MSA), enabling information exchange across window boundaries.

The computational burden of conventional global multi-head self-attention scales as expressed in Eq. (1.1):

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C \quad (1.1)$$

In contrast, the windowed variant achieves the complexity defined in Eq. (1.2):

$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC \quad (1.2)$$

Here, h and w represent feature map spatial dimensions measured in patches, C denotes the channel dimensionality, and M specifies the window extent (set to 7 in standard configurations). The critical distinction lies in complexity scaling: global attention grows quadratically with spatial resolution $(hw)^2$, whereas windowed attention scales linearly since M remains constant regardless of input size.

This investigation evaluates three Swin Transformer configurations differing in capacity: Swin-Base containing approximately 88 million learnable parameters, Swin-Small with 50 million parameters, and Swin-Tiny comprising 29 million parameters. The standard Vision Transformer (ViT-B/16) serves as a baseline for comparative analysis [20].

1.3.4 Loss Function

Network optimization employs categorical cross-entropy as the training objective, quantifying divergence between predicted class distributions and ground-truth labels, as defined in Eq. (1.3):

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C y_{i,j} \cdot \log(p_{i,j}) \quad (1.3)$$

In this formulation, N represents the batch cardinality, C equals the number of target categories (three for this application), $y_{i,j}$ indicates whether sample i belongs to class j (binary indicator), and $p_{i,j}$ denotes the network's predicted probability assigning sample i to class j .

1.3.5 Training Configuration

All networks initialize from weights pretrained on ImageNet, leveraging learned visual representations before task-specific adaptation. Parameter updates follow the AdamW optimization algorithm incorporating L2 regularization with coefficient 1×10^{-5} . Learning rate initialization occurs at 1×10^{-4} , subsequently modulated by cosine annealing with warm restarts (restart period $T_0 = 10$ epochs, period multiplier $T_{mult} = 2$). Training terminates after 30 epochs maximum, with early cessation triggered if validation performance stagnates for 7 consecutive epochs. Mini-batch size remains fixed at 32 samples throughout training.

1.3.6 Evaluation Metrics

Quantitative assessment employs multiple complementary metrics. Overall accuracy measures the proportion of correctly classified samples. Precision quantifies the fraction of positive predictions that prove correct, while recall captures the fraction of actual positives successfully identified. The F1-score synthesizes precision and recall through their harmonic mean, as given in Eq. (1.4):

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1.4)$$

Component metrics are defined as $\text{Precision} = TP/(TP + FP)$ and $\text{Recall} = TP/(TP + FN)$, where TP , FP , and FN denote true positive, false positive, and false negative counts respectively. Class-wise metrics undergo macro-averaging to produce aggregate performance indicators. Discrimination capability is further assessed through Area Under the ROC Curve (AUC-ROC) and Average Precision (AP) derived from precision-recall analysis.

1.3.7 Explainability Framework

Model interpretability is addressed through Gradient-weighted Class Activation Mapping (Grad-CAM), which identifies image regions most influential to classification outcomes. This technique backpropagates gradients from the target class score to intermediate convolutional feature maps, computing importance weights for each spatial location. The resulting activation maps highlight discriminative regions, enabling verification that predictions rely on agriculturally meaningful visual features rather than spurious correlations.

1.4 RESULTS AND DISCUSSION

1.4.1 Model Performance Comparison

Table 1.1 presents the comparative performance of evaluated vision transformer models on the test set.

Table 1.1: Performance Comparison of Vision Transformer Models.

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)
Swin-Small	98.90	98.90	98.90	98.90
Swin-Base	95.60	95.60	95.60	95.60
Swin-Tiny	95.60	95.60	95.60	95.60
ViT-Base	93.41	93.41	93.41	93.41

All metrics reported as macro-averaged values.

The Swin Transformer Small achieved the highest test accuracy of 98.90%, outperforming both larger (Swin-Base) and smaller (Swin-Tiny) variants. This finding suggests that the Small variant provides an optimal balance between model capacity and generalization for this dataset scale. The ViT-Base model achieved 93.41% accuracy, demonstrating competitive but lower performance compared to Swin variants.

1.4.2 Per-Class Performance Analysis

Table 1.2 presents per-class F1-scores across all transformer models.

The Swin-Small model demonstrated exceptional class-wise performance with perfect identification of Healthy leaves (100% F1) and near-perfect detection of Miner (98.3%) and Weevil (98.4%) damage. Notably, the Miner class presented the greatest classification chal-

Table 1.2: Per-Class F1-Score (%) by Model.

Model	Healthy	Miner	Weevil
Swin-Small	100.0	98.3	98.4
Swin-Base	98.3	93.1	95.4
Swin-Tiny	100.0	93.1	93.8
ViT-Base	96.8	89.3	93.8

lence across all models, likely due to subtle visual characteristics distinguishing miner damage from other conditions.

1.4.3 Confusion Matrix Analysis

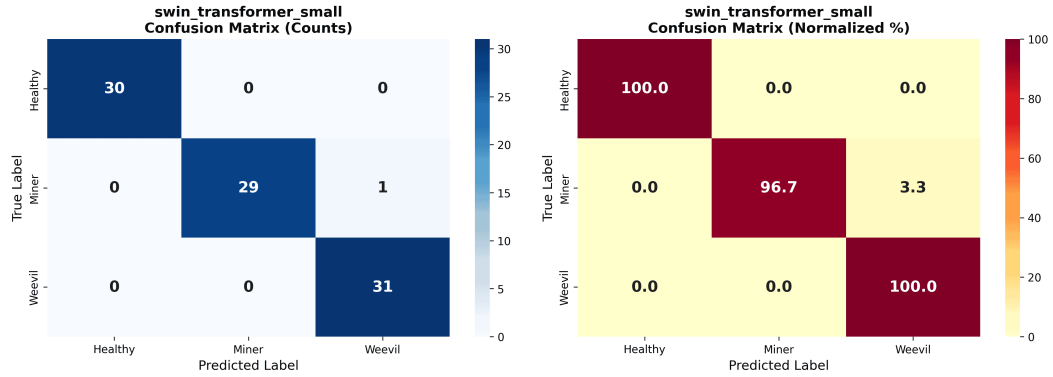


Figure 1.2: Confusion matrices for Swin Transformer Small.

Fig. 1.2 presents the confusion matrices for Swin Transformer Small on the test set in two formats: raw counts (left) and normalized percentages (right). The count matrix reveals the following classification performance:

- **Healthy:** 30/30 correctly classified (100% recall)
- **Miner:** 29/30 correctly classified (96.7% recall), with 1 sample misclassified as Weevil
- **Weevil:** 31/31 correctly classified (100% recall)

The normalized matrix confirms perfect classification for Healthy and Weevil classes at 100%, with Miner achieving 96.7% recall. The single misclassification involved a Miner-damaged sample predicted as Weevil-damaged (3.3%), suggesting potential visual similarity in certain damage patterns. Notably, zero false positives occurred for the Healthy class, indicating reliable identification of undamaged leaves—critical for avoiding unnecessary pesticide application in precision agriculture contexts.

1.4.4 ROC and Precision-Recall Analysis

The ROC curves demonstrated perfect discrimination capability with AUC values of 1.000 for all three classes. Similarly, Precision-Recall curves exhibited Average Precision (AP) of 1.000 across classes. These metrics indicate that the Swin-Small model achieves optimal separation between class probability distributions, enabling confident classification at various threshold settings.

1.4.5 Training Dynamics

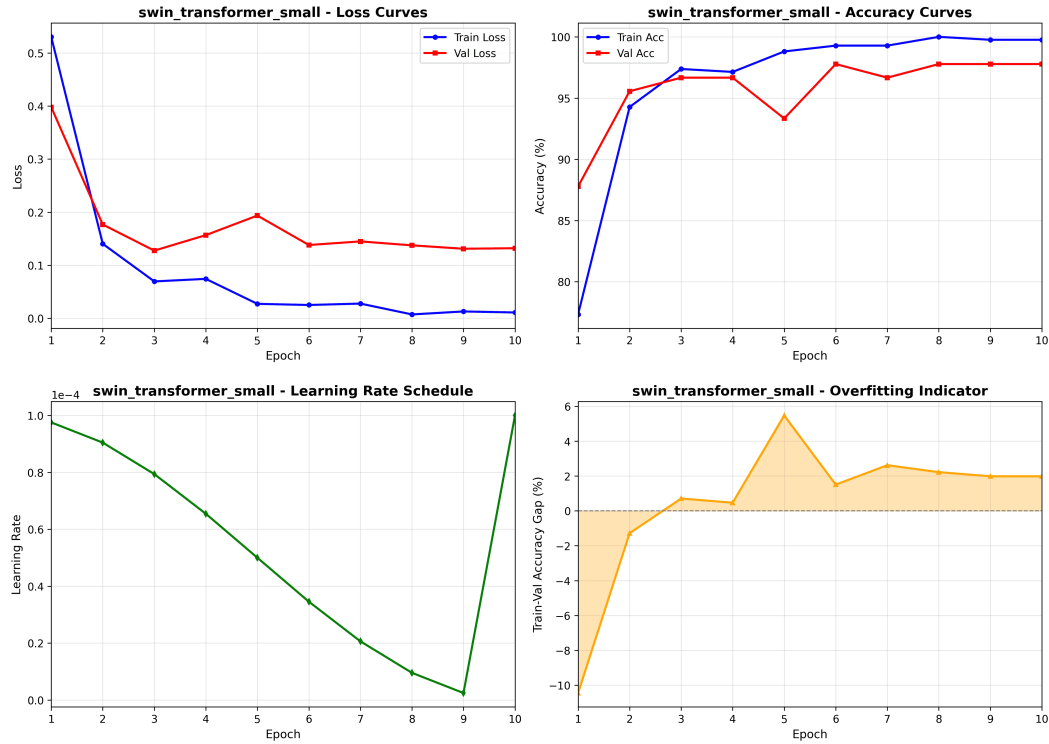


Figure 1.3: Training dynamics of Swin Transformer Small.

Fig. 1.3 presents the training dynamics of Swin Transformer Small over 10 epochs through four subplots. The loss curves (top-left) demonstrate rapid convergence with training loss decreasing from 0.52 to approximately 0.01, while validation loss stabilized around 0.13 after epoch 3. The accuracy curves (top-right) show training accuracy reaching nearly 100% and validation accuracy stabilizing at 97–98% after epoch 6. The learning rate schedule (bottom-left) follows cosine annealing, gradually decreasing from 1×10^{-4} to near-zero by epoch 9 before restarting. The overfitting indicator (bottom-right) displays the train-validation accuracy gap, which remained below 6% throughout training, indicating controlled generalization despite the model's capacity. The initial negative gap in early epochs suggests the model benefited from pretrained weights, with validation performance temporarily exceeding training performance before stabilizing.

1.4.6 Explainability Analysis

To validate that the model bases predictions on biologically meaningful features, Grad-CAM visualizations were generated for representative samples from each class. Each visualization comprises six panels: the original image with prediction confidence, Grad-CAM overlay on the predicted class, attention heatmap, and class-specific Grad-CAM activations for Healthy, Miner, and Weevil categories.

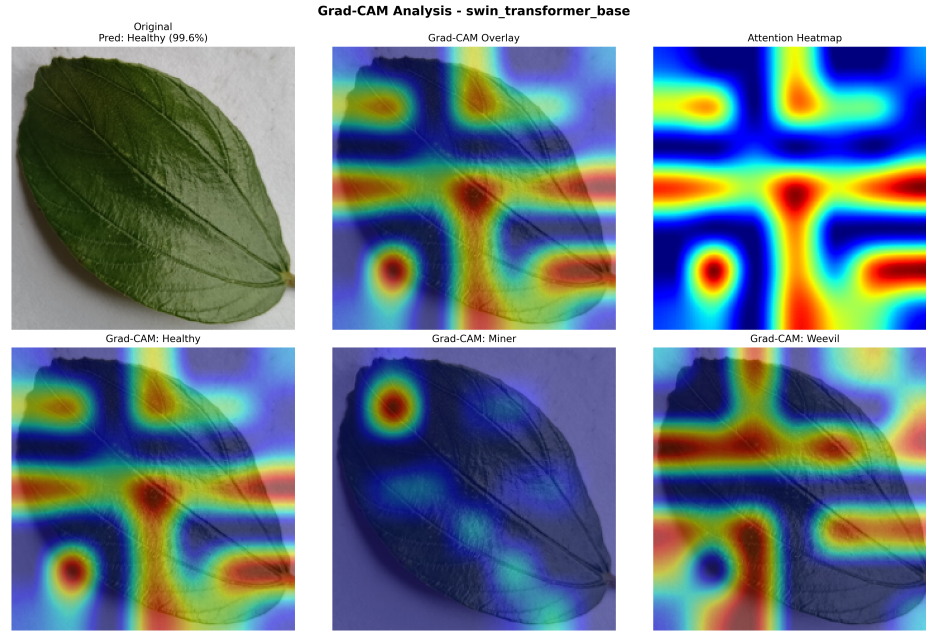


Figure 1.4: Grad-CAM visualization for Healthy leaf classification.

Fig. 1.4 presents the Grad-CAM analysis for a Healthy leaf sample, correctly classified with 99.6% confidence. The attention heatmap exhibits a distributed pattern across the leaf surface, indicating the model evaluates overall leaf integrity rather than focusing on localized regions. The class-specific visualizations show that Healthy activation (bottom-left) responds to uniform leaf texture, while Miner and Weevil activations remain suppressed in areas lacking damage characteristics.

Fig. 1.5 illustrates the Grad-CAM output for a Miner-damaged leaf, predicted with 100% confidence. The Grad-CAM overlay prominently highlights the serpentine mining trails and discolored tissue regions characteristic of leaf miner feeding patterns. The attention heatmap concentrates activation on the central damaged area, confirming the model identifies the distinctive tunneling damage caused by miner larvae. The Miner-specific activation (bottom-center) shows strong response to the affected regions while other class activations remain minimal.

Fig. 1.6 displays the analysis for a Weevil-damaged leaf, classified with 100% confidence. The model correctly focuses on the characteristic hole created by weevil feeding, as evidenced by the concentrated activation around the perforation in the Grad-CAM overlay. The attention heatmap shows strong response to the damaged region and surrounding tissue. The

Weevil-specific activation (bottom-right) demonstrates highest intensity at the feeding hole, validating that the model learned to recognize the distinctive circular damage pattern typical of weevil infestation.

These visualizations confirm that the Swin Transformer model attends to pest-specific damage patterns consistent with agronomic expertise, rather than spurious correlations or background features. This interpretability is essential for deploying the system in practical agricultural settings where stakeholder trust requires transparent decision-making.

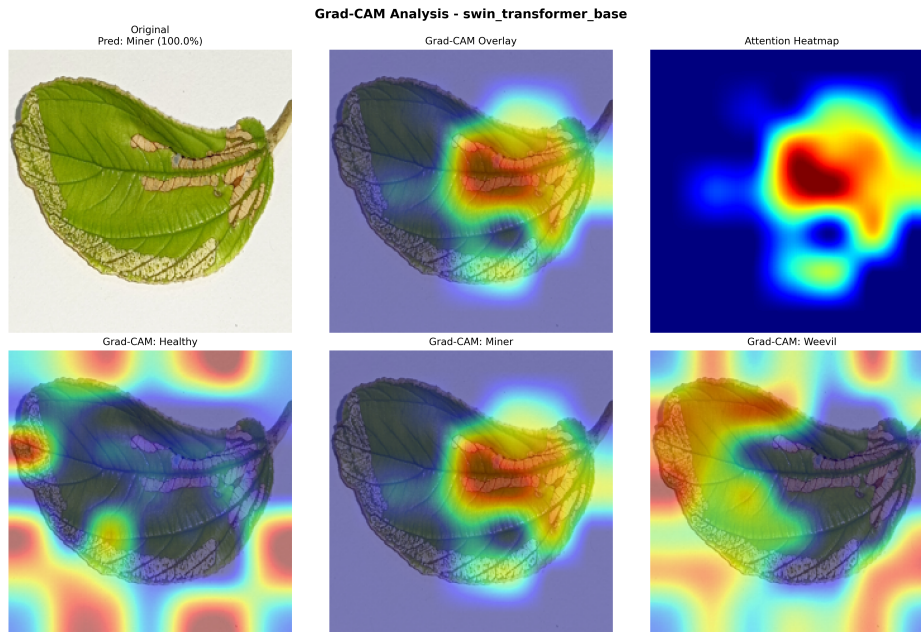


Figure 1.5: Grad-CAM visualization for Miner damage classification.

1.4.7 Architectural Insights

The superior performance of Swin-Small over Swin-Base can be attributed to several factors. First, the smaller parameter count (50M vs 88M) provides appropriate model complexity for the dataset scale avoiding overfitting while maintaining representational capacity. Second, the hierarchical shifted window mechanism enables effective capture of both local damage patterns and global leaf context without quadratic complexity. Third, ImageNet pretraining provides robust low-level feature representations that transfer well to agricultural imagery.

The performance gap between Swin variants and ViT-Base (5.49% accuracy difference) underscores the advantages of hierarchical processing for fine-grained agricultural image classification. While ViT's global attention captures long-range dependencies, it may overlook localized damage patterns that Swin's windowed attention effectively identifies.

1.4.8 Practical Implications

The achieved 98.90% accuracy presents significant practical value for precision agriculture:

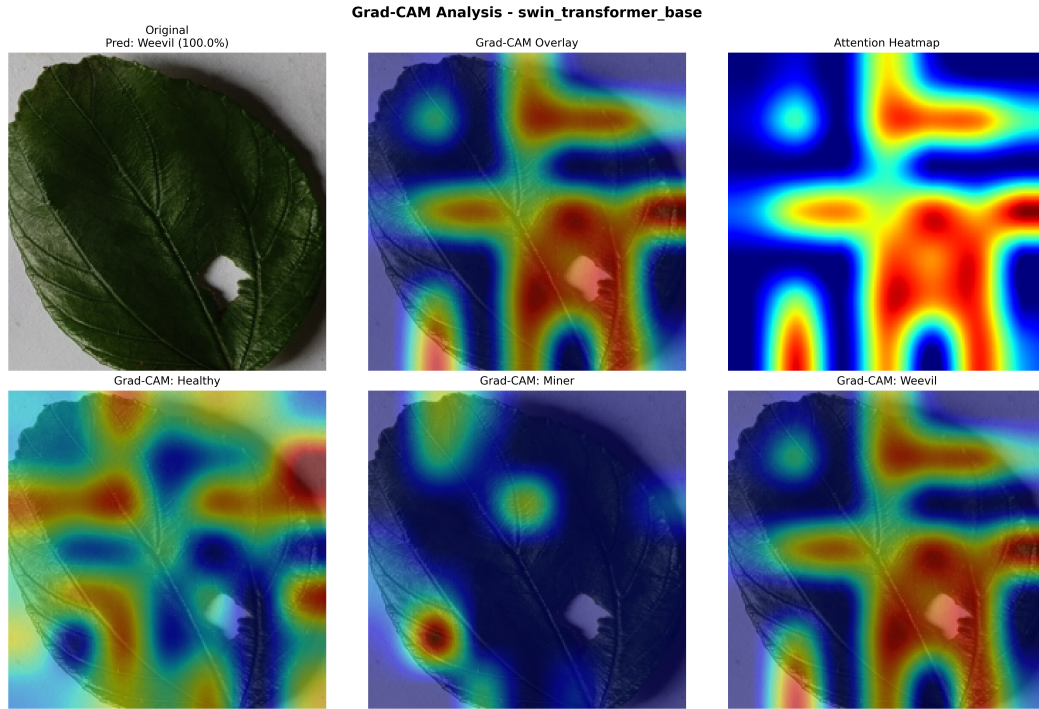


Figure 1.6: Grad-CAM visualization for Weevil damage classification.

- **Early Detection:** Automated identification enables timely intervention before pest populations reach damaging thresholds.
- **Targeted Treatment:** Accurate pest type classification supports selection of appropriate pest-specific control measures, reducing broad-spectrum pesticide usage.
- **Scalability:** The computational efficiency of Swin Transformer enables deployment on edge devices for real-time field monitoring.
- **Cost Reduction:** Automated monitoring reduces labor requirements while improving detection consistency compared to manual inspection.

1.5 CONCLUSION

This chapter demonstrates the effectiveness of Swin Transformer architecture for automated jujube leaf pest damage classification. The Swin Transformer Small variant achieved optimal performance with 98.90% test accuracy, outperforming both larger Swin variants and standard Vision Transformer. The model exhibited perfect AUC-ROC scores (1.000) and near-perfect F1-scores across all three classes (Healthy: 100%, Miner: 98.3%, Weevil: 98.4%), with only a single misclassification between Miner and Weevil damage categories.

The shifted window self-attention mechanism proved particularly effective for capturing both localized damage patterns and global leaf context, while maintaining computational

efficiency suitable for practical deployment. Transfer learning from ImageNet pretrained weights combined with appropriate data augmentation contributed to robust generalization despite limited agricultural training data.

The integration of explainable AI techniques ensures model interpretability, addressing the critical requirement for trustworthy AI systems in agricultural decision-making. Future work will expand the dataset to include additional pest types and damage severity levels, explore lightweight model variants for mobile deployment, and conduct field validation studies to assess real-world performance under diverse environmental conditions. The proposed framework establishes a foundation for intelligent pest management systems supporting sustainable jujube cultivation through precision agriculture technologies.

Bibliography

- [1] Li, Q., Yin, Z., Tan, W., Sun, X., Cao, H. & Wang, D. The resistance of the jujube (*Ziziphus jujuba*) to the devastating insect pest *Apolygus lucorum* (Hemiptera, Insecta) involves the jasmonic acid signaling pathway. *Pestic. Biochem. Physiol.* **196** art. 105597 (2023).
- [2] Sarwar, M. Incidence of Insect Pests on Ber (*Zizyphus jujube*) Tree. *Pak. J. Zool.* **38** pp. 261–263 (2006).
- [3] Kumar, M. R., Kandukuri, P., Srinivas, V. S., Mukkapati, N. & Syma Kumar, D. N. V. AI-SCAN: Advancing plant leaf disease detection with Transcendental Residual Convolutional Swin Transformer with hybrid optimizer. *CABI Agric. Biosci.* **6** art. 22 (2025).
- [4] Shahriyar, M. F. & Chy, A. M. R. Fine-Tuning Pretrained CNNs for Potato Tuber Disease Classification Using Soft and Hard Voting Ensembles. *IEEE 4th International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*. (2025).
- [5] Shahriyar, M. F., Das Gupta, S., Afjal, M. I., Datta, T. & Das Gupta, D. PTuberXplain: Addressing Class-Imbalanced Potato Tuber Disease Classification via Deep Transfer Features, Feature Selection and Explainable AI. *2026 IEEE International Conference on Electro Information Technology (eIT)*. (2026).
- [6] Ejaz, M. H., Bilal, M., Habib, U., Attique, M. & Chung, T.-S. Crop Pest Classification Using Deep Learning Techniques: A Review. *arXiv preprint arXiv:2507.01494* (2025).
- [7] Sharma, N., Al-Yarimi, F. A. M., Bharany, S., Rehman, A. U. & Taye, B. M. Hybrid AI Model With CNNs and Vision Transformers for Precision Pest Classification in Crops. *Food Sci. Nutr.* **13(11)** art. e71174 (2025).
- [8] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*. pp. 10012–10022 (2021).

- [9] Ahmed, M. T., Ahmed, M. W. & Kamruzzaman, M. A Systematic Review of Explainable Artificial Intelligence for Spectroscopic Agricultural Quality Assessment. *Comput. Electron. Agric.* **235** art. 110354 (2025).
- [10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Proc. IEEE Int. Conf. Comput. Vision*. pp. 618–626 (2017).
- [11] Venkateswara, S. M. & Padmanabhan, J. Deep learning based agricultural pest monitoring and classification. *Sci. Rep.* **15** art. 8684 (2025).
- [12] Alzahrani, A. S., Iqbal, A., Zafar, W. & Husnain, G. Swin-YOLO-SAM: A hybrid Transformer-based framework for plant disease detection. *Front. Plant Sci.* **16** art. 1666374 (2025).
- [13] Yu, S., Xie, L. & Dai, L. ST-CFI: Swin Transformer with convolutional feature interactions for identifying plant diseases. *Sci. Rep.* **15** art. 25000 (2025).
- [14] Liu, W. & Zhang, A. Plant Disease Detection Algorithm Based on Efficient Swin Transformer. *Comput. Mater. Contin.* **82(2)** pp. 3045–3068 (2025).
- [15] Wu, Y., Li, X., Zhang, Q., Zhou, X., Qiu, H. & Wang, P. Recognition of spider mite infestations in jujube trees based on spectral-spatial clustering of hyperspectral images from UAVs. *Front. Plant Sci.* **14** art. 1078676 (2023).
- [16] Erdinç, A. & Erdoğan, H. Detection of the Metcalfa pruinosa (Hemiptera: Flatidae) pest on the Jujube plant (*Ziziphus jujuba*) using a sequence of YOLOv5 models. *Turkish J. Agric. Nat. Sci.* **11(3)** pp. 797–806 (2024).
- [17] Kondaveeti, H. K. & Simhadri, C. G. Evaluation of deep learning models using explainable AI with qualitative and quantitative analysis for rice leaf disease detection. *Sci. Rep.* **15** art. 31850 (2025).
- [18] Al Sahili, Z. & Awad, M. The power of transfer learning in agricultural applications: AgriNet. *Front. Plant Sci.* **13** art. 992700 (2022).
- [19] Md Mafiul Hasan, M., Oishe, S. E. & Ahmed, S. Jujube Diseases Identification Image Dataset. *Mendeley Data*, V2, doi: 10.17632/6vttpjr7xw.2 (2025).
- [20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proc. Int. Conf. Learn. Representations (ICLR)* (2021).