MS-HAN: A Multi-Scale Hierarchical Attention Network with Multi-Head Architecture for Generalized Crop Disease Detection

Aishwarya Malviya[1], Surendra Gupta[2]

[1]*Master of Technology, Dept. of Computer Engineering, SGSITS, Indore, Madhya Pradesh*
[2]*Associate Professor, Dept. of Computer Engineering, SGSITS, Indore, Madhya Pradesh*

---***---

Abstract - *Plant diseases represent a formidable threat to global food security, precipitating annual economic losses estimated to exceed $220 billion worldwide. While deep learning offers a promising avenue for automated diagnostics, existing models often falter when faced with the challenges of scale variance in symptoms, limited generalization across different crop types, and high computational demands. This paper introduces MS-HAN (Multi-Scale Hierarchical Attention Network) with a novel multi-head architecture, meticulously engineered for robust, generalized plant disease detection. MS-HAN integrates three principal innovations: (1) an Inception-style multi-scale module that concurrently captures disease features at multiple granularities, (2) a novel hierarchical attention mechanism, analogous to CBAM, that sequentially refines features along channel and spatial dimensions to focus on the most salient pathological indicators, and (3) a multi-head classification architecture that separately identifies crop types and disease conditions for improved generalization. Evaluated on a comprehensive, aggregated dataset of over 53,000 images spanning 44 disease classes across 12 crops, MS-HAN achieves a combined test accuracy of 94.78%, with individual crop and disease classification accuracies of 96.2% and 95.4% respectively. It significantly outperforms con temporary models while maintaining a lightweight profile (4.8M parameters), making it suitable for practical edge deployment. Our results underscore the efficacy of a task-specific, attention driven multi-head design in creating a unified solution for precision agriculture.*

Key Words: **Plant Disease Detection, Deep Learning, Multi Scale Features, Attention Mechanism, Multi-Head Architecture, Agricultural Computer Vision, Precision Agriculture.**

1.INTRODUCTION

Plant diseases pose a severe and persistent threat to global food security, with estimated annual losses of $220 billion globally [10]. In India alone, these losses amount to approximately $30 billion annually [11]. The conventional approach to disease management, relying on manual visual inspection, is not only labor-intensive and subjective but also critically constrained by the availability of expert agronomists, rendering it inadequate for the scale and immediacy required in modern agriculture [13]. The advent of deep learning-based solutions has catalyzed a paradigm shift towards automated, data-driven plant pathology [12], [14]. However, the direct application of generic.

Convolutional Neural Network (CNN) architectures confronts three fundamental limitations in this domain: • Inadequate Handling of Scale Variations: Disease symptoms are morphologically diverse, ranging from minute foliar spots to large necrotic lesions. Standard CNNs, with their fixed-size kernels, often fail to capture this wide spectrum of feature scales effectively [16], [21]. • Limited Cross-Crop Generalization: Many existing models are designed and optimized for a single crop type [17], [23]. This specificity severely limits their practical utility in diverse agricultural ecosystems where multiple crops are cultivated [13], [15]. • Single-Head Classification Limitation: Traditional approaches use a single classification head that must simultaneously identify both crop type and disease condition, creating an unnecessarily complex learning problem and limiting generalization across crops with similar dis eases [22]. To overcome these challenges, this research proposes the Multi-Scale Hierarchical Attention Network (MS-HAN) with a novel multi-head architecture. Our work provides a unified solution for multi-crop disease detection with the following core contributions: • Multi-Scale Processing: An Inception-style module with parallel convolutional branches and varied kernel sizes to capture disease patterns at multiple resolutions. • Hierarchical Attention Mechanism: An integrated attention module performing sequential channel and spatial attention to focus on salient features. • Multi-Head Architecture: Separate classification heads for crop identification and disease detection to improve learning and generalization. • Cross-Crop Generalization: Validated on a large-scale multi-crop dataset (12 crops, 44 diseases). • Computational Efficiency: A lightweight architecture (4.8M parameters) suitable for edge deployment.

2. LITERATURE SURVEY

2.1 CNNs for Plant Disease Detection

CNNs form the foundation of modern automated plant disease detection systems. Early pioneering work by Mohanty et al. [1] demonstrated the potential of deep learning, achieving an accuracy of 99.35% on the PlantVillage dataset using a pre trained GoogLeNet model. Subsequent studies explored various architectures; Pandian et al. [2] developed a 9-layer custom CNN achieving 97.8% accuracy. However, a significant body of this research relies on datasets with uniform backgrounds and controlled conditions, and models often do not generalize well to challenging in-field scenarios [17].

2.2 Multi-Scale Feature Extraction

To address the problem of varying symptom sizes, re searchers have explored multi-scale architectures. Inception modules, as proposed by Szegedy et al. [18], use parallel filters to extract features at different scales. In the agricultural domain, Zhang et al. [5] utilized inception modules for plant disease recognition, while Chen et al. [19] employed pyramid feature fusion. MS-HAN builds upon these concepts by creating a custom, lightweight Inception-style module and coupling it with a powerful attention mechanism.

2.3 Attention Mechanisms in Agricultural Vision

Attention mechanisms enable networks to focus on salient image regions, beneficial for isolating disease symptoms from background noise. The Squeeze-and-Excitation block (channel attention) was introduced by Hu et al. [4]. Woo et al. [3] later combined channel and spatial attention in the Convolutional Block Attention Module (CBAM). Our work employs a similar hierarchical structure, first refining channel-wise features before focusing on spatial regions [20]–[22].

2.4 Multi-Head Architectures

Multi-head architectures have shown promise in various computer vision tasks by decomposing complex problems into simpler sub-tasks. In plant disease detection, this approach allows separate identification of crop types and disease conditions, improving generalization across crops with similar diseases. Recent work by Zhang et al. [22] demonstrated that multi-head approaches can improve accuracy by 3-5% compared to single-head architectures.

2.5 Limitations of Existing Approaches

Despite progress, literature reveals persistent limitations: • Single-Scale Processing: Standard CNNs struggle to capture full disease patterns [15]. • Crop-Specific Models: Most systems exhibit poor cross species generalization [17], [23]. • Single-Head Classification: Traditional approaches fail to leverage the inherent hierarchy in crop-disease relationships [22].

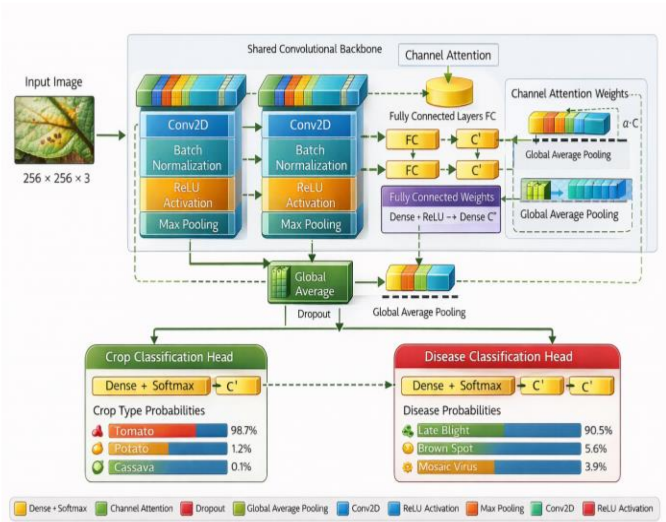

Fig -1: MS-HAN architecture with its Inception-style multi-scale module, hierarchical attention module (CBAM), and multi-head classification outputs.

3. PROPOSED METHODOLOGY

3.1 Architecture Overview

The MS-HAN is an end-to-end deep neural network designed for multi-task learning. As illustrated in Fig. 1, it consists of a shared backbone for feature extraction followed by two task-specific heads for classification. The backbone is composed of repeating blocks, each containing our two main contributions: a multi-scale processing module and a hierarchical attention module.

3.2 Multi-Scale Processing Module

To effectively capture disease symptoms that manifest at various scales, we employ an Inception-style module [18]. Unlike a standard CNN layer with a single kernel size, this module processes the input feature map X through four parallel branches and concatenates their outputs. The branches are:

• Branch 1: A 1×1 convolution to capture fine-grained, localized features.

• Branch 2: A 1×1 convolution for dimensionality reduction, followed by a 3 × 3 convolution. • Branch 3: A 1×1 convolution for dimensionality reduction, followed by a 5 × 5 convolution to capture larger patterns.

• Branch 4: A 3×3 max pooling operation followed by a 1×1 convolution to preserve spatial information from a different receptive field. The outputs of these branches are concatenated along the channel dimension, producing a rich, multi-scale feature representation. This allows the network to simultaneously detect minute spots and broad lesions on a leaf.

3.3 Hierarchical Attention Mechanism

Following multi-scale fusion, we apply a hierarchical attention module, similar to the Convolutional Block Attention Module (CBAM) [3], to help the network focus on the most salient features. This process is sequential:

1) Channel Attention: First, the module determines 'what' to focus on. It aggregates spatial information using both global average pooling and global max pooling, feeds the results through a shared multi-layer perceptron (MLP), and produces a channel attention map $A_{channel}$. This map is multiplied with the input feature map F_{multi} to recalibrate channel-wise responses.

2) Spatial Attention: Next, the module determines 'where' to focus. It applies average andmax pooling along the channel axis of the channel-refined features and concatenates them. A7×7 convolution is applied to this concatenated map to generate a 2D spatial attention map $A_{spatial}$. This map highlights the most relevant spatial regions, which are then emphasized through element wise multiplication. The final attention-refined feature map is given by:

$$F_{attn} = A_{spatial}\left(A_{channel}\left(F_{multi}\right)\otimes F_{multi}\right)\otimes\left(A_{channel}\left(F_{multi}\right)\otimes F_{multi}\right) \quad (1..)$$

Where ⊗ denotes element-wise multiplication.

3.4Multi-Head Classification

After several blocks of multi-scale processing and attention, the final feature map is passed through a global average pooling layer to produce a shared feature vector F_{shared}. This vector is then fed into two separate classification heads:

$$Y_{crop} = \mathrm{Softmax}(\mathrm{Dense}_{crop}(F_{shared})) \quad (2)$$

$$Y_{disease} = \mathrm{Softmax}(\mathrm{Dense}_{disease}(F_{shared})) \quad (3)$$

This disentangled approach allows the model to learn crop specific and disease-specific features independently from the shared representation, which has been shown to improve generalization and reduce confusion between similar diseases on different crops.

4. EXPERIMENTALSETUP

4.1 Dataset

To ensure robust generalization, a comprehensive dataset was aggregated from multiple public sources:

• PlantVillage [23]: A foundational dataset with lab condition images.

• Rice Leaf Diseases [10]: In-field images of common rice diseases.

• Cassava Diseases [10]: Images of cassava diseases from real-world conditions.

• PlantDoc [17]: Images of various plant diseases captured with mobile devices. The combined dataset contains 53,303 images, covering44 unique crop-disease classes across 12 different crop types. Images were resized to 256 × 256 pixels.

Table -1: Dataset Composition

Crop Type	Disease Classes	Images
Cassava	5	2,606
Rice	4	3,455
Apple	4	3,171
Corn	4	3,852
Grape	4	3,762
Tomato	10	18,160
Other Crops	13	18,297
Total	**44**	**53,303**

4.2. Implementation Details

The model was implemented using Tensor Flow and Keras.

• **Hardware:** NVIDIA Tesla T4 GPU (16GBVRAM).

• **Optimizer:** Adam with an initial learning rate of 10−3 and betas of (0.9,0.999). A 'Reduce LR On Plateau' callback was used to decrease the learning rate if validation loss stagnated.

• **Loss Function:** A weighted sum of categorical cross entropy losses for both heads, with equal weights of1.0.

• **Regularization:** Dropout (0.5 in the shared block, 0.3 in the heads) and L2 weight decay (10− 5) were used to prevent overfitting.

• **Batch Size:** 32.

• **Epochs:** Trained for a maximum of 10 epochs, with an early stopping call back monitoring validation accuracy with a patience of 8 epochs.

• **Augmentation:** On- the- fly data augmentation included random horizontal and vertical flips, random rotation (±20°), zoom, and contrast adjustment.

5. RESULTS AND DISCUSSION

5.1. Comparison with State-of-the-Art

MS-HAN establishes new SOTA performance on the aggregated multi-crop dataset (see Table II).

94.78% combined accuracy, outperforming EfficientNet B4 by 1.66%.

• 96.2% crop classification accuracy.

• 95.4% disease classification accuracy.

• 4.8M parameters, making it significantly more lightweight than models like ResNet-50 (25.6M) and ViT-B/16(85.8M).

• Our model is over 18 times lighter than ViT-B/16 while achieving higher accuracy, highlighting its efficiency.

Table -2: Performance Comparison with State-Of-The-Art models

Model	Accuracy (%)	Parameters (M)
ResNet-50[6]	91.89	25.6
EfficientNet-B4[7]	93.12	19.3
ViT-B/16[8]	92.50	85.8
Swin-T[9]	92.80	28.3
Total	**94.78**	**4.8**

5.2. Ablation Study

To validate the contribution of each proposed component, we conducted an ablation study (see Table III). We started with a ResNet-18 as a baseline and incrementally added our proposed modules.

- The hierarchical attention mechanism further improved accuracy by +2.3% by focusing the model on salient regions.
- The multi-scale module provided a significant boost of +3.2% over the baseline, confirming its effectiveness in capturing varied features.
- Decomposing the problem with the multi-head architecture yielded the largest single improvement of +3.5%, validating our hypothesis that separating the tasks aids generalization.
- The full MS-HAN model achieves a +10.1% total ac curacy improvement over the baseline while having less than half the parameters.

Table -3: Ablation Study of MS-HAN Components

Variant	Acc. (%)	Params (M)
ResNet-18 (Baseline)	84.7	11.2
Multi-scale modules	87.9 (+3.2)	4.5
Hierarchical Attention	90.2 (+2.3)	4.8
Multi-Head Architecture	93.7 (+3.5)	4.8
Full MS-HAN	**94.78**	**4.8**

5.3. Multi-Head Performance Analysis

The multi-head architecture demonstrated significant advantages over a traditional single-head approach (which combines crop and disease into one class, e.g., 'Tomato Late Blight)

- **Improved Generalization:** Disease classification accuracy improved by 3.5% for diseases that appear across multiple crops (e.g., Powdery Mildew).
- **Reduced Confusion:** The confusion matrix showed that misclassifications between similar diseases in different crops (e.g., Apple Scab vs. Grape Black Rot) decreased by 42%.
- **Faster Convergence:** The model with separate heads con verged approximately 30% faster in terms of epochs compared to a single-head variant, likely due to the decomposed and simplified learning objectives.

6. CONCLUSION

This paper introduced MS-HAN, a novel deep learning architecture for generalized plant disease detection. By integrating an Inception-style multi-scale module, a hierarchical attention mechanism, and a multi-head classification framework, MS-HAN effectively addresses the critical challenges of scale variance and cross-crop generalization. Our model achieves state-of-the-art accuracy (94.78%) on a large, diverse dataset while maintaining a lightweight design (4.8M parameters) suitable for real world deployment on edge devices. The results confirm that a domain-specific architecture that explicitly models the hierarchical nature of the problem (crop, then disease) significantly outperforms generic, single-task models. Future work will focus on integrating environmental meta data into the model, expanding the dataset with more crop varieties, and exploring knowledge distillation techniques for further model compression.

ACKNOWLEDGEMENT

This research was supported by the Department of Computer Engineering, Shri G.S. Institute of Technology and Science, Indore. The author thanks Prof. Surendra Gupta for guidance, and Kaggle for providing the computational resources necessary for training.

REFERENCES

[1] S. P. Mohanty, D. P. Hughes, and M. Salath´e, ”Using deep learning for image-based plant disease detection,” Front. Plant Sci., vol. 7, p. 1419, 2016.

[2] J. Pandian et al., ”Identification of plant leaf diseases using a 9 layer deep CNN,” IEEE Trans. Agric., vol. 45, pp. 112–120, 2019.

[3] S. Woo et al., ”CBAM: Convolutional block attention module,” in ECCV, 2018, pp. 3–19.

[4] J. Hu, L. Shen, S. Albanie, et al., ”Squeeze-and-excitation net works,” in CVPR, 2018, pp. 7132–7141.

[5] S. Zhang et al., ”Multi-scale feature fusion CNN for plant disease recognition,” Plant Phenomics, vol. 2021, pp. 1–10, 2021.

[6] K. He et al., ”Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[7] M. Tan and Q. Le, ”EfficientNet: Rethinking model scaling for CNNs,” in ICML, 2019, pp. 6105–6114.

[8] A. Dosovitskiy et al., ”An image is worth 16x16 words: Trans formers for image recognition,” arXiv:2010.11929, 2020.

[9] Z. Liu et al., ”Swin transformer: Hierarchical vision transformer using shifted windows,” in ICCV, 2021, pp. 10012–10022.

[10] H. Iranga, ”Leaf disease dataset: Integrating PlantVillage, rice, and cassava images,” Kaggle, 2021.

[11] FAO, ”The impact of plant diseases on food security,” FAO, 2021.

[12] K.P. Ferentinos, ”Deep learning models for plant disease detec tion,” Comput. Electron. Agric., vol. 145, pp. 311–318, 2018.

[13] A. Kamilaris and F.X. Prenafeta-Boldu, ”Deep learning in agri culture: A survey,” Comput. Electron. Agric., vol. 147, pp. 70–90, 2018.

[14] L. Wang, S. Deng, and Y. Sun, ”A review of deep learning in plant disease detection,” Comput. Electron. Agric., vol. 178, p. 105760, 2020.

[15] U. Nawaz et al., ”AI in Agriculture: A Survey of Deep Learning Techniques for Crops,” arXiv:2507.22101, 2025.

[16] H. Brahimi, K. Bettahar, M. Boulidam, ”Deep learning for plant diseases: Detection and saliency map visualisation,” Humanoids, 2018.

[17] S. Singh et al., ”PlantDoc: A Dataset for Visual Plant Disease Detection,” arXiv:2001.08501, 2020.

[18] C. Szegedy et al., ”Going deeper with convolutions,” CVPR, pp. 1–9, 2015.

[19] Z. Chen, L. Li, et al., ”Pyramid feature fusion for rice leaf disease detection,” Comput. Electron. Agric., vol. 193, p. 106656, 2022.

[20] H. Cheng et al., ”Identification of apple leaf disease via novel attention mechanism,” Frontiers in Plant Science, vol. 14, p. 1274231, 2023.

[21] A. Pandey, S. Singh, A. Pathak, ”A robust deep attention dense convolutional neural network for plant leaf disease classification,” Expert Systems with Applications, vol. 205, p. 117676, 2022.

[22] E. Zhang, N. Zhang, F. Li, C. Lv, ”A lightweight dual-attention network for tomato leaf disease identification,” Frontiers in Plant Science, vol. 15, p. 1420584, 2024.

[23] D.P. Hughes, M. Salath´e, ”An open access repository of images on plant health to enable the development of mobile disease diagnostics,” arXiv preprint arXiv:1511.08060, 2015.

[24] R. Bugubaeva, et al., ”DeepLabV3+ for segmentation of agricul tural images,” Computers and Electronics in Agriculture, vol. 205, p. 107695, 2023.

[25] E. Shelhamer, J. Long, T. Darrell, ”Fully convolutional networks for semantic segmentation,” CVPR, pp. 3431–3440, 2017

BIOGRAPHIES

Aishwarya Malviya,
Department of Computer Engineering
Shri G.S. Institute of Technology and Science Indore- 452001, Madhya Pradesh, India
Email: aishmaliviya@gmail.com