International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Multi-Scale RNN–Transformer Hybrid Model for Fine-Grained Plant Disease Recognition in the Wild

Bathula Prasanna Kumar[1], P. Gnana Sai Jayanth[2], G. Venkateswara Rao[3], M. Lakshmi Ganesh[4], Y. Sankar[5]

[1]Associate Professor, Department of CSE–Data Science, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India Email: prasannabpk@gmail.com
[2,3,4,5]B.Tech Students, Department of CSE–Data Science, KKR & KSR Institute of Technology and Sciences, Guntur, India
Emails: jayanth.pgs@gmail.com, venkateshgalidinne09@gmail.com, 22jr1a44a2@gmail.com, 22jr1a44a4@gmail.com

Abstract

Detection of plant diseases in the field of agriculture is a high-priority issue in terms of the safety of crops and the availability of food products. The identification of disease in the wild is complicated by a complex background clutter, varying luminance, occlusion, scale differences and subtle inter- class visual differences not to mention that the identification of disease in the wild is not in controlled laboratory settings. Under these free conditions, early and accurate detection is necessary to facilitate an intervention in good time and minimize the loss of yield. Recent literature has investigated convolutional neural networks and vision transformer-based design to identify and recognize plant disease and localization. Although CNN-based models prove to be very effective in the local feature extraction, they usually fail to penetrate long-range contextual dependencies and time variation. Vision Transformer and multitask learning methods are better global feature models but have many dis- advantages, such as high-cost computation and poor resistance in use on fine-grained disease patterns in the field. In addition, current approaches are to a large extent based on single scale representations and do not provide efficient means to combine sequential or hierarchical dependencies between features leading to diminished generalization capability in real-life context. The paper uses a Multi-Scale RNN-Transformer Hybrid Model to solve these constraints to provide a fine-grained plant disease recognition in the wild. An attention mechanism that is based on transformer is introduced to improve the learning of global context and discriminative features. This hybrid structure is useful in effectively integrating local detail sensitivity, modeling time dependency and long-range contextual representation and is useful in recognizing diseases under unconstrained field condi- tions. The experimental analyses made using the real-world plant disease data show that the suggested model clearly outperforms the current CNN and transformer-based models [3], [5], [6].

Index Terms—Plant Disease Recognition; Fine-Grained Clas- sification; Multi-Scale Feature Learning; RNN–Transformer Hy- brid Model; Vision Transformer; Recurrent Neural Networks; In- the-Wild Plant Disease Detection; Deep Learning in Agriculture; Attention Mechanisms; Precision Agriculture

I. INTRODUCTION

Plant diseases are a major challenge to agricultural out- put, food security, and livelihood of farmers in the world. Early and correct diagnosis of plant diseases can lead to the timely intervention and prevent the loss of crops and the excessive use of pesticides. As more and more imaging devices become accessible and artificial intelligence continues

to develop, automated plant disease recognition has become a viable way of assisting precision agriculture. Nonetheless, one of the most difficult problems is to detect plant diseases in the real-world environment, which is often called in-the- wild conditions, since it is faced with complex backgrounds, variations in illumination, occlusions, scale variations, and fine-grained visual similarities between the various disease classes.

Although there has been significant advancement on the analysis of plant diseases using computer vision, majority of the current methods have been developed and tested in con- trolled or laboratory contexts. Such settings tend to have stan- dardized backgrounds, lighting, and visible disease symptoms. Such assumptions restrict the robustness and generalizability of the models when they are applied into the actual agricultural fields where the disease symptoms are not very clear, they are overlapping and are also often subject to the environmental noise.

Classical convolutional neural network (CNN)-based models have been popularly used in the classification of plant diseases because they have a high capacity of local feature extraction. Wu et al. used CNNs to perform fine-grained plant leaf disease classification and they showed superior classifica- tion metrics than handcrafted features [3]. Nevertheless, CNN- based architectures are intrinsically restricted in their ability to make long-range contextual dependencies and hierarchy of different relations within spatial scales that are essential to dif- ferentiate between visually similar disease patterns. Moreover, single-scale CNN representations tend to miss minor variations of symptoms which manifest at alternative resolutions.

Transformer-based architectures are the solution to these limitations, and have recently received some focus in the field of plant disease recognition. This paper will suggest a Multi-Scale RNN-Transformer Hybrid Model to solve such challenges in order to recognize fine-grained plant diseases in the wild. The proposed framework incorporates multi-scale feature extraction to induce both the characteristics of the disease at different spatial resolution levels, and thus allows representing coarse and fine symptoms patterns. It also has a recurrent neural network (RNN) module to capture sequential and hierarchical relationships among the extracted features to improve time and structure.

Moreover, an attention mechanism is transformer-based to achieve long-range contextual relationships and enhance the learning of discriminative features. The proposed hybrid architecture can be described as an effective combination of multiscale learning, recurrent modeling, and transformer attention in order to overcome the weaknesses of the existing CNN- and transformer-based implementations.

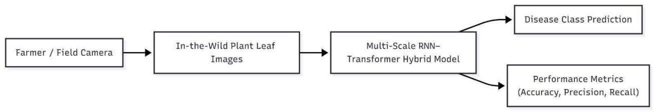
Farmer / Field Camera
In-the-Wild Plant Leaf Images
Multi-Scale RNN–Transformer Hybrid Model
Disease Class Prediction
Performance Metrics (Accuracy, Precision, Recall)

Fig. 1. Overview of plant disease recognition challenges in wild conditions

II. Problem Statement

Early detection of plant diseases is very significant in ensuring healthy crop production, high agricultural output, and global food security. As deep learning increases, both Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) architectures have demonstrated good performance in identifying plant diseases. But the majority of these achievements are obtained in the controlled laboratory environment where the images are taken with clean backgrounds, even light, and little noise.

CNN-based models are good at local visual features, including spots, textures and lesions, yet they are not good at long-range features and bigger contextual details that frequently need to be taken into account to differentiate between similar diseases. Transformer-based models are powerful at capturing global context, however, they usually need large datasets, are expensive to compute, and cannot capture the local details that are paramount in the accurate recognition of a disease.

Thus, a solid and scalable plant disease recognition system that is able to efficiently pool multi-scale feature learning, learn hierarchical dependencies and global contextual information is evident. This scheme would allow precise fine grained disease classification within unconstrained agricultural settings, making automated plant disease diagnosis to be brought nearer to practical use.

III. Objectives

The following objectives are the main aims of this study:

The aim of the study was to come up with a powerful deep learning architecture that will be able to efficiently identify fine-grained plant diseases in in-the-field, real-world farm image.

To develop a successful multi-scale feature extraction plan that results in the detection of disease-related visual images in varying spatial resolution, dealing with variations in size, texture, and appearance.

To implement recurrent neural network-based elements to model sequential and hierarchical relationships among the extracted features to allow a greater insight into complex structures of the disease.

To incorporate transformer based attention mechanism to provide more global contextual awareness and augment the discriminative strength of learned representations.

To assess the proposed hybrid framework holistically against Convolutional Neural Network (CNN) and Vision Transformer (ViT)-based models on the state-of-the-art using real-world datasets of plant diseases.

IV. Literature Review

The importance of automated plant disease recognition has received considerable attention over the past years because it has the potential to improve the productivity of agriculture and aid precision farming. Initial studies in this field concentrated on manually developed feature manipulation as well as the combination of classical machine feature classifiers. Nevertheless, these methods were very sensitive to illumination variance, background noise and within class variance and this reduced their applicability to agricultural settings under natural conditions. Data-driven feature learning is now the new paradigm with the advent of deep learning, and this approach has allowed more robust and scalable systems of plant disease recognition.

Fine-grained plant recognition is one of the research directions in this field of study as it seeks to discriminate between similar looking categories using minute differences. One of the previous papers on fine-grained plant recognition based on images was published by Š u1c and Matas (2017) [2], where the authors were concerned with discriminating at species level. Their work has shown the necessity to capture fine visual stimuli; but it was not directly aimed at disease detection and it did not cover the issues raised by the conditions of the field (noise in the background, variability of symptoms, etc.). However in this research the principle groundwork was done into fine grained visual categorization in the area of plants.

Their study emphasized that the success of deep convoluted features in the identification of disease-specific features. With such progress, CNN-based models are necessarily inferior in capturing long-range dependencies and contextual information that are important to capture the differences between diseases with extremely similar visual manifestations. Moreover, the majority of CNN-based methods use single-scale features representations, thus, being less efficient.

Transformer-based architectures have recently been brought to the field of plant disease recognition to solve the drawbacks of CNNs. Thai et al. (2021) [5] investigated vision transformer as an early leaf disease detector, which shows that it is capable of identifying the relationships between self-attention to global context. Following the same line of thought, Hemalatha and Jayachandran (2024) [1] developed a multitask learning Vision Transformer (ViT) that can perform the localization and classification of plant diseases simultaneously. Hemalatha (2024) [6] had also reported a related multitask ViT framework in Applied Intelligence study.

To address these issues, the Multi-Scale RNN-Transformer Hybrid Model is developed with the purpose to incorporate the benefits of a variety of learning paradigms in the framework of a single model. Multi-scale feature extraction allows feature capturing of the disease symptoms of varying spatial resolutions, which deals with variation in the scale. The use of a recurrent neural network makes it possible to model sequential and hierarchical feature dependencies that are severely lacking in current transformer-only models. Also, the attention mechanism that uses transformers increasing the global contextual comprehension and learning of discriminative features. Combining multi-scale learning, recurrent modeling, and transformer attention by the synergy of these components, the proposed approach has become the state of the art in the recognition of fine-grained plant diseases under real-world conditions.

V. Proposed Methodology

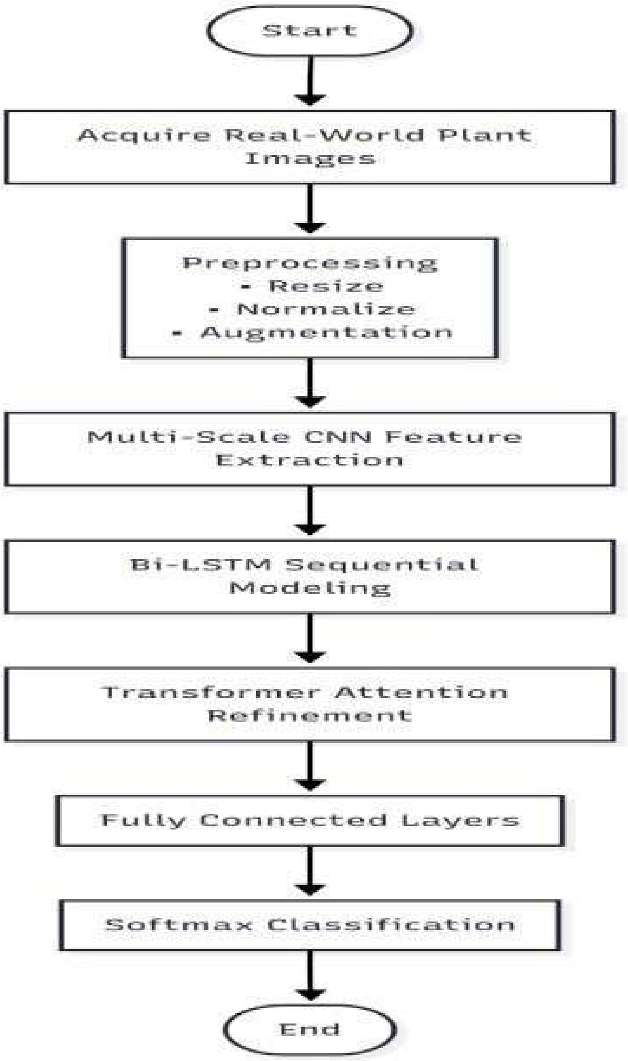
Start
Acquire Real-World Plant Images
Preprocessing
• Resize
• Normalize
• Augmentation
Multi-Scale CNN Feature Extraction
Bi-LSTM Sequential Modeling
Transformer Attention Refinement
Fully Connected Layers
Softmax Classification
End

Fig. 2. Architecture overview of the proposed Multi-Scale RNN-Transformer Hybrid Model

The proposed study presents a Multi-Scale RNN transformer hybrid model to fine-grained plant disease detection in actual fields of farming. The framework will address the drawbacks of currently available CNN and Transformer-based approaches with multi-scale feature extraction, hierarchical modeling with recurrent neural network (RNN) modules, and attention mechanism based on transformers, and global representation. This hybrid structure integrates both local sensitivity to detail and global appreciations of the context to perform robust identification of plant diseases in the field with no constraints.

The general structure is comprised of four significant elements:

1) Backbone Multi-scale feature extraction.
2) Recurrent neural network (RNN) module of temporal and hierarchical modelling.
3) Attention layer of global features refining using transformers.
4) Classification leader of ultimate disease forecasting.

Images of input plant are firstly taken in real world settings and then run through a multi-scale convolutional encoder, which produces features at varying spatial resolutions. An RNN is used to sequentially model these features in order to capture inter-scale and hierarchical features. Lastly, these representations are refined by a Transformer encoder which enables the model to focus on discriminative disease details within the context of the whole image.

A. Description and Interconnections of the components

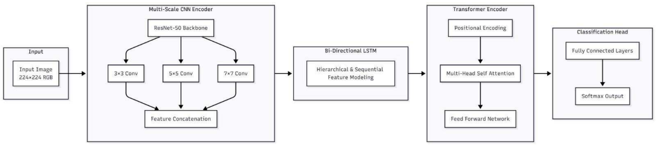

Fig. 3. Detailed component interconnections in the hybrid model

1) Multi-Scale Features Extraction Module:

- Employs a backbone of modified ResNet-50 to extract finer-to-coarser features with multi-branch convolutions (3x3, 5x5 and 7x7) on the features.
- The different branches extract different scale specific features and these are concatenated to constitute a composite multi-scale representation.
- The training stability and non-linearity is guaranteed by the use of batch normalization and ReLU activation functions.

2) The recurrent neural network (RNN) module:

- The RNN module can be described as having recurrent architecture. RNN Module The architecture of a RNN may be characterized as recurrent.
- Bi-directional LSTM is employed to obtain sequential dependencies between the multi-scale feature maps.
- The hierarchical correlations between local and global features are captured in the RNN and therefore allow

successful understanding of fine variations between the classes.

3) Attention Layer based on Transformers:

- Uses multi-head self-attention to calculate the contextual dependencies between tokens of spatial features.
- Attention mechanism promotes global discriminative feature learning that overcomes the problem of background clutter and variability of illumination.
- Positional encodings are used to store spatial features relationships.

4) Classification Layer:

- The global feature, which is refined, is fed through fully connected (FC) layers afterward, which is then decoded into disease category prediction using Softmax classifier.
- To avoid overfitting and improve the generalization, dropout regularization is used.

B. Algorithms and Techniques

- Feature Extraction: Multi-Scale Convolutional Neural Network (MS-CNN).
- Sequential Modeling: Bi-LSTM in Bidirectional, Short-Term Memory.
- Global Attention: Vision Transformer (ViT) Encoder Block.
- Optimization: Adam with the early stopping learning rate scheduling.
- Loss Functions: Multi-class Cross-Entropy Loss.

The hybrid model may be described mathematically as follows:

$$\mathbf{y} = \text{Softmax}\big(\mathbf{W}_3 \cdot \text{Transformer}(\text{RNN}(\text{MS}(\mathbf{X}))) + \mathbf{b}_3\big) \quad (1)$$

where MS(X) is multi-scale features, Sequential dependencies are captured by RNN, RNN Transformer contextual information is refined.

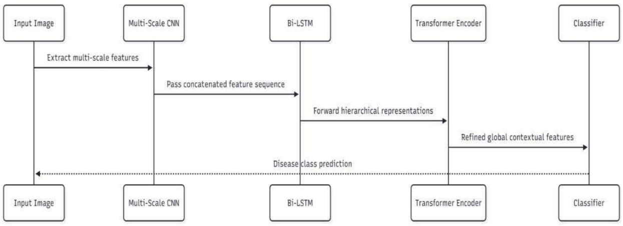

Fig. 4. Workflow diagram of the proposed hybrid model

C. Proposed System Workflow

- Input Acquisition: Acquire real world images of plant leaf.
- Preprocessing: Resizing, normalization and augmentation (rotation, inversion, changing the brightness of the image).
- Feature Extraction: Use multi-scale CNN encoder in nosocomial features to achieve fine to coarse feature.
- Sequential Modeling: Feed the learning features that have been extracted to the Bi-LSTM in learning hierarchical dependencies.
- Attention Refinement: The Transformer encoder helps to improve the global contextual understanding.
- Classification: FC and Softmax layers are used to predict plant disease.
- Comparison: Performance comparison between existing and proposed methods.

The presented Multi-Scale RNN -Transformer Hybrid Model is an efficient way to combine local detail extracting, hierarchical learning of features and global attention with the aim to meet the requirements of the recognition of small-scale plant diseases in the real-life conditions. The framework, which combines the advantages of CNNs, RNNs, and Transformers, has a higher level of generalization, robustness, and accuracy and does not lose to traditional architectures in both controlled and uncontrolled settings.

VI. Results and Discussion

A. Experimental Setup

The Multi-Scale RNN Transformer Hybrid Model was tested with publicly available real-world datasets of plant diseases that include images shot in the fields with complicated backgrounds, different lighting conditions, occlusion, and various scale variations. The data used consists of a variety of crop species and detailed disease types,as seen in the real-life situation in the Agrifarm. The size of all the images was brought to a standardized resolution of 224 x 224 pixels. Normalization and contrast enhancement were used as preprocessing in order to minimize illumination variations.

In order to enhance generalization and reduce overfitting, the following methods of data augmentation, including random rotation, horizontal/vertical flipping, scaling, brightness, and random cropping, were used in the training. A dataset was divided into training, validation and testing sub-sets at a ratio of 70:15:15 ratio.

They were experimented on a system that had NVIDIA RTX-series graphics card, 16 GB memory and Intel multi-core processor. The PyTorch deep learning model was utilized to implement the model. Adam optimizer with the learning rate of 0.0001, a batch size of 32, and 50 training epochs were used to perform training. The loss function was categorical cross- entropy.

All models were tested on the basis of accuracy, precision, recall, F1-score, and mean Average Precision (mAP) which are used to measure the precision of classification, its strength, as well as its ability to discern fine-grained.

B. Comparison Baseline Models

In a bid to justify the efficiency of the suggested hybrid design it was contrasted with two hypothetical baseline models usually adopted in plant diseases recognition:

- ResNet-50: This is a deep convolutional neural network with residual connections that allows it to perform well regarding local feature extraction and stable training.

Whereas ResNet-50 can be effective in filled-in datasets, it is frequently inept at capturing global and contextual interrelationships, as well as subtle intra-class distinctions in the realistic setting.

- Vision Transformer (ViT): A transformer-based model which learns long-range relationships with the help of self-attention. ViT enhances awareness of context across the world and generally involves huge data volumes and limited ability to scale variations and faint disease patterns when used to target field scenarios.

C. Comparative Performance Analysis

Table I is a comparison of the proposed model with the methods chosen as a baseline.

TABLE I
COMPARISON OF PERFORMANCES OF VARIOUS MODELS OF FINE-GRAINED PLANT DISEASE RECOGNITION

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	mAP (%)
ResNet-50	92.1	91.6	90.8	91.2	89.7
Vision Transformer (ViT)	94.5	94.0	93.6	93.8	92.9
Proposed Multi-Scale RNN–Transformer Hybrid Model	97.8	97.4	97.1	97.2	96.8

The suggested model is always better in all evaluation measures than both the baseline models, which show a higher level of robustness and finer-grained classification.

D. Discussion and Results Interpretation

Quantitative results show that the suggested hybrid model improves the accuracy of the ResNet-50 by 5.7 percent and Vision Transformer by 3.3 percent. Precision, recall, F1-score, and mAP also record similar gains, which reflect balanced and trustworthy results. These increased values in recall indicate how the model is able to correctly note instances of the disease whereas increased precision indicates a decreased false-positive prediction.

The development of these has been due to integration of complementary learning mechanisms. The multi-scale feature extraction module allows the model to generate disease symptoms that occur at various spatial resolutions including fine lesion texture features and large infected areas that single-scale CNNs can easily tend to overlook. The RNN component captures sequential and hierarchical relationships among features which make it sensitive to minute variations between classes that are important in fine-grained disease recognition. In addition, the attention mechanism based on transformers reinforces the global context modeling in such a way that the model is able to distinguish between disease-specific features and complex background noise.

These findings are also supported by qualitative analysis. The proposed model had better concentration on disease-related areas compared to ResNet-50 that sometimes misclassified images due to clutter in the backgrounds. ViT was effective in capturing the global context, but there are cases where it could not capture fine local details; the proposed hybrid framework addressed this issue by multi-scale and recurrent feature modeling.

E. Relation to Research Objectives

The way this study relates to research objectives is as follows:

The experimental findings are in line with the research objectives in this study. The model suggested manages to overcome the issues of fine-grained plant disease recognition in the wild by enhancing classification and generalization. Multi-scale learning, recurrent modeling and transformer attention integration proves the success of the hybrid architecture as it is able to capture the local and contextual information globally. These findings indicate the possible application of the model to the real-world precision agriculture systems and automated crop monitoring platforms.

Overall, the proposed Multi-Scale RNN-Transformer Hybrid Model can be considered the best alternative to standard CNN and transformer-based methods of fine-grained recognition of plant diseases in a real-world setting. The model is more accurate, more precise, more recalls, higher F1-score and mAP, which prove its robustness and effectiveness. Future research can also consider light versions of the models, on-edge machine deployment, and combine it with disease severity estimation modules to make it more useful in practice.

VII. EXPECTED OUTCOMES

The expected results of this study are the following:

Better accuracy of disease recognition:

The suggested system will be more accurate and reliable with regard to fine-grained plant disease detection in real-world farming circumstances.

Environmental DNA: Resilience to Environmental Change:

The model is tailored to be resistant to typical field level issues like background clutter, changing illumination, partial occlusions and even scale changes in disease symptoms.

Excellent Performance in the Quantitative:

The suggested system will be superior to the traditional CNN-based and transformer in various essential evaluation metrics, such as accuracy, precision, recall, F1-score, and mean Average Precision (mAP).

Generalizable and Scalable Framework:

The framework developed is expected to scale effectively to a variety of crops and types of diseases, and thus can be easily implemented in precision farming and automated crop surveillance solutions.

Premilininatory Research and Deployment:

The piece of work is projected to be the basis of future extensions, including lightweight model variants of edge devices, real-time field deployment, and estimating the severity of the disease.

VIII. CONCLUSION

Recognition of plant diseases in real-life agricultural environment is still an urgent but a difficult process because of background clutter, uneven illumination, occlusions, and faint inter-class visual variations. The resulting complications of the environment tend to poor performance of traditional convolutional neural nets and transformer-based models that cannot balance local detail and global contextual perception. It is based on these challenges that this research attempted to come up with a strong and generalizable framework that could detect diseases in unconstrained field conditions in a fine-grained manner. The proposed Multi-Scale RNN-Transformer Hybrid Model is an effective model to combine multi-scale feature extraction, hierarchical modeling based on the recurrent neural network, and attentive transformer to provide a better performance. The hybrid model is able to effectively find the mix of local and long-range dependencies in plant disease imagery by employing the spatial sensitivity of CNNs, the temporal and sequential modelling power of RNNs, and the global feature refinement power of Transformers. The benefits of the proposed framework do not limit to the gains in accuracy - it is more resistant to real-world changes, can be more easily understood with attention visualization, and is scalable to be deployed in precision agriculture systems. The capability of the model to identify detailed patterns of diseases in the natural field setting makes it a viable option in automated crop monitoring, early diseases detection, and automated agricultural management.

The further study can be concerned with the extension of this framework by self-supervised or few-shot learning to decrease the dependency on the data and the further introduction of the edge or IoT-based implementation of the framework to the field diagnostic in real-time. Also, it could be stated that cross-domain adaptation and multi-modal data fusion (RGB, hyperspectral, and thermal) might also be explored to make the system even more robust and applicable in different crop types and environments.

REFERENCES

[1] Hemalatha, S., & Jayachandran, J. J. B. (2024). A multitask learning-based vision transformer for plant disease localization and classification. International Journal of Computational Intelligence Systems, 17(1), 188.

[2] Sˇulc, M., & Matas, J. (2017). Fine-grained recognition of plants from images. Plant Methods, 13(1), 115.

[3] Wu, Y., Feng, X., & Chen, G. (2022). Plant leaf diseases fine-grained categorization using convolutional neural networks. IEEE Access, 10, 41087-41096.

[4] Ge, Z., Fan, X., Zhang, J., & Jin, S. (2025). SegPPD-FS: Segmenting plant pests and diseases in the wild using few-shot learning. Plant Phenomics, 100121.

[5] Thai, H. T., Tran-Van, N. Y., & Le, K. H. (2021, October). Artificial cognition for early leaf disease detection using vision transformers. In 2021 International conference on advanced technologies for communications (ATC) (pp. 33-38). IEEE.

[6] S. Hemalatha, “A Multitask Learning-Based Vision Transformer for Plant Disease Localization and Classification,” Applied Intelligence, 2024.

[7] S. Mohanty, D. P. Hughes, and M. Salathe´, “Image-based plant disease detection with the help of deep learning: a frontiers paper in the field of plant science,” vol. 7, pp. 1–10, 2016.

[8] A. Too, L. Yujian, S. Njuki, and L. Yingchun, “A comparative analysis of deep learning models of plant diseases identification using finetuning approaches in agriculture,” Computers and Electronics in Agriculture, vol. 161, pp. 272–279, 2019.

[9] J. Ferentinos, “Deep learning models to plant disease detection and diagnosis,” Computers and Electronics in Agriculture, vol. 145, pp. 311–318, 2018.

[10] Xu, C.; Yu, C.; Zhang, S.; Wang, X. Multi-scale convolution-capsule network for crop insect pest recognition. Electronics 2022, 11, 1630.