Intelligent Multimodal Framework for Explainable Plant Disease Diagnosis and Treatment Recommendation

Ayesha Butalia*[1], Yash Nimbalkar[2], Reena Gunjan[3]

[1] Professor, School of Computing, MIT-ADT University, Pune, India
[2] Student, School of Computing, MIT-ADT University, Pune, India
[3] Professor, School of Computing, MIT-ADT University, Pune, India
*Corresponding author ayesha.butalia@mituniversity.edu.in, yaashnimbalkar@gmail.com

Abstract. Plant diseases continue to pose a significant challenge to food security in the world and this has resulted into extensive loss in crop yield and financial instability among agricultural communities. The conventional approaches to the disease diagnosis are based on manual examination by specialists and are thus time-consuming, subjective, and can never be done at large-scale levels. Most of the existing systems, even recent developments that have achieved significant progress in automated plant disease recognition using leaf image as the single-modality visual input, cannot operate with contextual reasoning, interpretability, and adaptable decision-making, although recent progress in deep learning, especially Convolutional Neural Networks (CNNs) and YOLO-based architectures, has substantially increased the accuracy and effectiveness of automated plant disease recognition. In order to overcome these drawbacks, this paper presents an Agentic AI Framework, which combines visual crop image analysis, natural language symptom description, and structured agricultural knowledge to provide explainable and reliable plant disease diagnosis. The model makes use of coordinated autonomous agents that perform the functions of vision perception, language understanding, multimodal fusion, retrieval-augmented reasoning, and treatment planning. The system is capable of predicting the diseases with accuracy and giving evidence-based and practical treatment recommendations by using the authoritative agronomic resources of ICAR and FAO. Evidence-based on experimental validation of the proposed framework on the basis of PlantVillage and PlantDoc databases proves a higher accuracy of the diagnostics, as well as greater trust, making the given framework a comprehensive decision-support system of sustainable and reasonable crop management.

Keywords: Agentic AI, Multimodal Reasoning, Explainable Diagnosis, Precision Agriculture, Knowledge Retrieval.

1 Introduction

Agriculture is a key pillar in the global food security and agricultural plant diseases have been known to cause significant losses in crop volume, quality and economic stability. In most cases, these ailments are contagious and thrive well in a healthy environment hence early and accurate diagnosis is crucial in preserving production systems. The traditional diagnostic procedures are founded on the visual inspection of the specialists, and this is time-consuming, labor intensive and have subjectivity

S. Bhalerao et al. (eds.), *Proceedings of the 2nd International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2026)*, Advances in Intelligent Systems Research 207,
https://doi.org/10.2991/978-94-6239-678-4_9

differences. In recent Deep learning models such as Convolutional Neural Networks (CNNs) have become highly effective in terms of their uses over the years good at diagnosing visual patterns of diseases with an image of a leaf [1]. Such models as YOLOv8 and YOLOv11 have optimized the feature extraction and enhanced the real time detection capabilities as well as localization accuracy [2,3]. Generative and diffusion-based models have contributed to this as well by adding benefits and improving datasets in terms of their performance under variable fields. Although this progress has remained limited as existing systems are predominantly based on single-modality image data and lack contextual reasoning, expert knowledge integration, and decision support capabilities, incompetence to reason and minimal integration of appropriate agricultural know-how. The events occurring recently highlight the potentialities of Agentic AI paradigms that utilize autonomy, memory and self-adaptation to support complex decision-making frameworks in constantly changing environments [4]. However, most of the existing implementations are restricted to the classification that is based on the visual, and they do not involve contextual information, such as a description of the symptoms, weather patterns, and organizational experience of ICAR and FAO. To eliminate these inadequacies, this study suggests an Agency Multimodal AI Framework that unites the interpretation of pictures, perception of symptoms in writing, and knowledge of agriculture, boosted with retrieval, into a decision system.

2 Literature Review

The concept of artificial intelligence has emerged as one of the major forces behind contemporary agricultural automation, especially in the area of disease detection of plants and monitoring of crops health, which is facilitated by recent progress in deep learning and agent-based intelligence [1]. Convolutional Neural Networks (CNNs) and similar image-based methods have shown excellent performance in disease-specific texture, color differences, and lesion patterns in images of leaves [2]. The current versions of the YOLO architecture have gone further to enhance real-time detection of blight, rust, and leaf-spots symptoms with realistic field conditions [3]. In order to address the constraint of datasets, Generative Adversarial Network (GAN) type augmentation techniques are considered to generate realistic leaf images and increase the robustness of the model, whereas Explainable Artificial Intelligence (XAI) techniques are used to visualize decision heatmaps and make these models more interpretable and trusted by users [4]. Nevertheless, image-only methods are also vulnerable to environmental noise and context differences. Consequently, there has been growing interest in multimodal systems incorporating both visual and weather information and soil conditions and textual descriptions of symptoms. Multimodal models based on diffusion have shown greater stability and generalization in a variety of conditions of agricultural fields [5].

Table 1. Performance Comparison with Existing Systems.

Problem	Existing approach	Limitation	Proposed solution
Real-time field detection	YOLO-based vision models	Vision-only; lacks context	Vision agent with multimodal fusion
Class imbalance	GAN and CNN based pipelines	Limited field generalization	Confidence-aware augmentation
Explainability	XAI overlays	No agronomic grounding	RAG over ICAR/FAO knowledge
Cross-crop generalization	Transfer learning	No treatment planning	Agent-based planner
System coordination	Early agentic concepts	Partial integration	End-to-end agentic pipeline

3 Methodology

3.1 Farmer Input and Data Profiling

The first step in the process is the gaining of the necessary inputs from the farmers, which include the image of the leaves, the symptoms of the plants in natural language, and the environment, which includes the temperature, humidity, and moisture level of the soil. This is the first pattern of the system, and this will ensure that the system is aware of the type of plant, the stage of growth, and the diseases that are prevalent in the area. Just like the context portraying in the intelligent wellness system, this will ensure that the system is conversant with the context of the area of application before engaging in the diagnosis, thereby ensuring better contextual accuracy of the further prediction [1]. The environment, which includes the temperature, humidity, and moisture level of the soil, is included in the system for better context awareness during the diagnosis.

3.2 System Design and Workflow

The proposed framework ensures that it follows a multi-agent paradigm where multiple agents are working in unison to produce a final diagnosis and advisory output. The image quality assessment is carried out, and the appropriate analysis of the image is determined by the Triage Agent. The Vision Agent identifies the disease using the best architectures such as CNNs and YOLOv11 to identify high-resolution visual features. The Text Agent uses language embeddings to interpret the qualitative observations of the farmer to process the symptoms. The Fusion Agent combines these multiple modal features to produce a single decision context. The Guideline Agent retrieves knowledge-base and retrieval augmented validated treatments using knowledge-base lookups and retrieval augmented techniques. The Planner Agent converts the disease identification and retrieved guidelines to a structured treatment plan.

3.3 Data Integration and Multimodal Analysis

After processing of individual data streams, the system uses a multimodal fusion mechanism to synthesize and analyze such streams together. Vision agent image embeddings and text agent text embeddings are mapped into a common feature space. This combination takes images, description of symptoms and other contextual aspects in one representation. The similarity of the fused embedding is compared to disease signature stored in a structured knowledge graph using cosine similarity. This enables model to do context-sensitive classification and thus the diagnosis is stronger in cases where the images are noisy, symptoms are similar and in cases where the environment has effects on the occurance of the disease [6].

3.4 Knowledge Retrieval and Generation of Treatment

After the diagnosis, the system switches on its knowledge-retrieval component. Guideline Agent retrieves the treatment protocol using Retrieval-Augmented Generation (RAG) to access the treatment protocols in the expert approved agricultural handbooks, ICAR/FAO documents, and other approved sources that have been indexed by the system [5]. Such retrieved guidelines consist of disease-specific management practice, prescribed pesticides or organic remedies, dosage intervals, and the preventive measures. The Planner Agent checks this information to be restructured into a simple farmer-understandable action plan to include treatment steps, safety directions, and follow-up nt intervals.

3.5 Feedback and Adaptive Learning

The system uses a dynamic feedback mechanism to aid the continuous improvement. Feedback Agent collects the post-treatment feedback of the farmer on whether the recommendation can help or not, whether the symptoms got better or the misdiagnosis. The feedback entries are stored in a monitored retraining pool, where it can optimize its similarity mappings, retrieval priorities, and manipulates the reasoning of the agents over time. The system can improve its accuracy in decision-making and diagnostic confidence over time based on the principles of reinforcement learning. As new cases and responses of the farmers are fed to the system, it converts to a self-renewing agricultural intelligence system that can adapt to seasonal variations and new diseases and region-specific developments [7].

4 Computing For PlantAegis

4.1 Adaptive Agent Confidence Reinforcement (AACR)

To quantify how much confidence an agent gains after incorporating feedback from previous predictions.

$$A_{reinforce} = \lambda_1 \; C_{base} + \lambda_2 \, F_{corr} + \lambda_3 T_{usr} \tag{1}$$

Where, $A_{reinforce}$ – Reinforced confidence of an agent after adaptation, C_{base} – Baseline confidence from the original prediction, F_{corr} – Correction ratio from feedback data., T_{usr} – Trust score provided by user validation and $\lambda_1,\lambda_2,\lambda_3$ – Tunable weights controlling influence balance.

4.2 Disease Severity Progress Index (DSPI)

Used to evaluate the rate at which disease spreads based on sequential image captures

$$DSPI = \left(\frac{A_{lesion},t2 - A_{lesion},t1}{A_{leaf}}\right) \times 100 \quad (2)$$

Where, $A_{lesion},t1$ $A_{lesion},t2$ – Lesion area at two obs times and A_{leaf} – Total leaf area.

4.3 Multimodal Feature Relevance Index (MFRI)

Measures the relative contribution of each modality (vision, text, environment) to the final fusion output

$$MFRI_m = \frac{w_m . I_m}{\sum_{i=1}^{n} w_m . I_i} \quad (3)$$

Where ,MFRI – Relevance index for modality, w_m – Weight assigned to modality m, I_m – Mutual information between modality m and final label.

5 Proposed Model For PlantAegis.

5.1 Intelligent Multimodal Diagnosis Engine

The architecture that is proposed is initiated by the diagnosis engine that has the ability to comprehend the visual and text inputs of the farmers. A Vision Agent works with the leaf images using the improved YOLOv11 model that has the ability to identify small features of the lesions, from the perspective of irregular lighting and visibility. At the same time, the Text Agent processes the text inputs of the symptoms that the farmer has described using the transformer-based semantic encoders that have the ability to derive the hints of the disease from the stories that the farmer has narrated in the text format. With the combination of the visual and text inputs, the system is able to perform the diagnosis in a much more accurate way, and the possibility of misclassification based on the differences in the perspectives and the presence of noise in the inputs is greatly minimized. The confidence scores that have been obtained from the agents are standardised and developed using the confidence reinforcement strategy that has been defined in Equation (1).

5.2 Context-Aware Confidence Enhancement Layer

After the initial diagnosis, the system takes into account various contextual aspects of the microclimate, status of the soil, and the crop growth stage to improve the accuracy of the prediction. The system employs a Fusion Agent to use 3 parameters, namely α, β, and γ, to calculate the weighted confidence score, to which visual, textual, and environmental data are adaptively assigned. This allows the system to adapt the weights of the decision based on the results, providing accurate results in real time, as has been achieved in earlier models of experimentation in multimodal agricultural data models [3],[4]. The weights of the visual, textual, and environmental data are determined based on the relative weights of each of these data modalities, as determined by the multimodal feature relevance formulation defined by Equation (3).

5.3 Knowledge Retrieval and Knowledge Treatment Planner Evidence Driven

The Guideline Agent, after final diagnosis, makes a call to a Retrieval-Augmented Generation pipeline to obtain disease-specific treatment options based on the agronomic sources that have been validated. The sources include ICAR crop protection guides and FAO disease management manuals[6]. The pesticide/organic remedy recommendations, application schedule, dosage limits, and precautions comprise the retrieved knowledge. This knowledge is transformed into a planned action plan, which can be followed step by step by the Planner Agent. The temporal variations of disease impression are totally tracked at various stages of the analysis, similar to the disease severity progression measure defined in Equation (2).

5.4 Adaptive Learning and Feedback Reinforcements

In order to ensure that there is a constant development of the model, currently, a Feedback Agent is responsible for collecting post-treatment feedback of farmers, whether the remedy works, does not work, or works with certain modifications. The interaction of these is stored in a feedback memory pool, which is used to update similarity mappings, tuning parameters, and model weights over time. The use of post-decision feedback allows for the constant development of agent reliability, including diag-nostic confidence over time, in accordance with the mechanism of adaptive confidence update introduced in Equation (1). The mechanism of reinforced structure is responsible for learning slowly to react to region-specific diseases, peculiarities, season, and changes in pathogens [7].

5.5 Cost and Resource Optimization Module

There is also the additional module of cost estimation, which considers the cost-effectiveness of the treatment plans. It makes predictions regarding the cost of the entire expenditure based on the dosage required, the area of the crop, and the domestic prices of the inputs in the market. This will allow the system to be able to prescribe measures that have a balance between the biological effect and cost. Cost-awareness is added to system, which early AI-driven crop advisory systems mostly lacked [8].

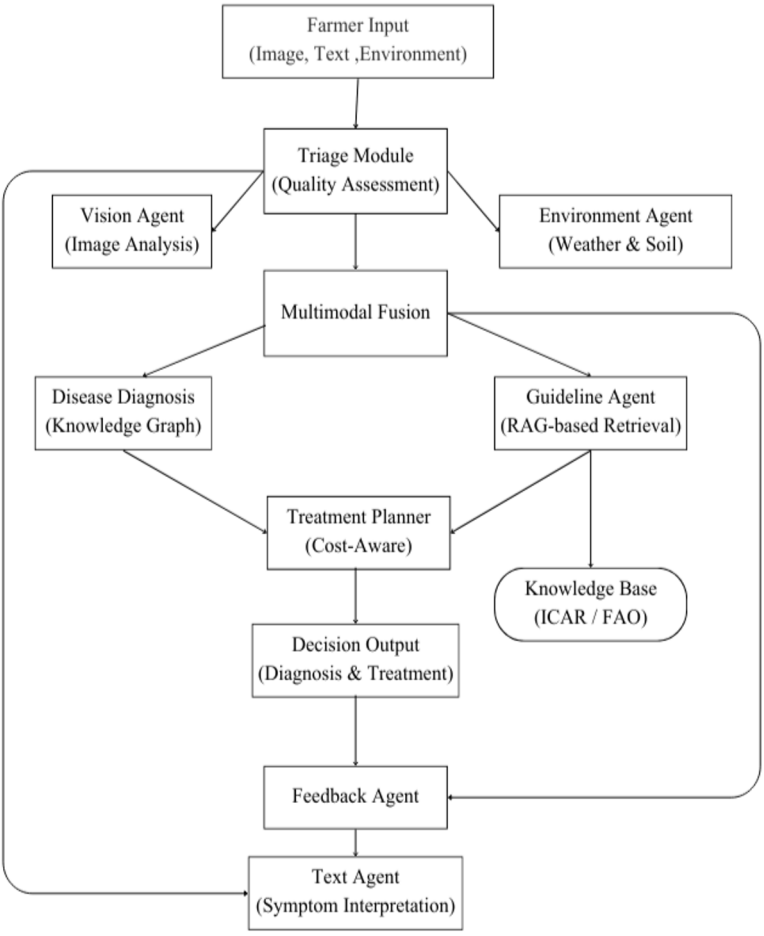

Fig. 1. Architecture of the Proposed PlantAegis Framework

6 Validation , Data Collection, Experimentation, Results and Interpretation.

6.1 Dataset

Knowledge and visual datasets are combined to make sure that good multimodal validation is provided. The leading training dataset is the PlantVillage dataset, which consists of 54,304 RGB leaf images of 14 crops with 38 disease types and healthy samples, which were obtained at Kaggle/UCI to train CNN, YOLOv8, and ViT models. In the experimental study of real performance, the contribution by the PlantDoc dataset on Mendeley contains 2,598 images, which have been collected on the field in 13 crops and 17 disease classes, adding to the variations of lighting, noise, and environmental conditions. The RAG knowledge base which serves as support to the Planner and Guideline Agents is comprised of known agricultural knowledge,

such as the approximately 4,200 structured entries on the pest disease links, region specific remedies, and crop management practices which are maintained by the ICAR and FAO. They are combined to form a balanced basis of enhancing the visual detection accuracy of the visual object and the text-based reasoning of the decision taken in the form of robust multimodal plant disease analysis.

6.2 Experimentation

The hybrid environment between TensorFlow and PyTorch was used to measure the diagnostic accuracy, interpretability, and robustness of the multimodal framework. PlantVillage and Plant Doc datasets were subjected to preprocessing and image normalization, augmentation, and Sentence-BERT embeddings of text inputs, and an 8020 train validation split was taken into consideration. A joint training of the Vision Agent (YOLOv8 and ViT) and Text Agent (SBERT on ICAR FAO symptom descriptions) and Fusion Agent with adaptive weigh produced multimodal results. This is also advanced by the Retrieval-Augmented Generation that has region-specific FAO and ICAR guidelines on pesticide safety and pesticide treatment recommendation. The performance of the system is assessed on the basis of accuracy, precision, recall, F1-score, Trust Score, and Cost Optimization Ratio, thus, confirming the fact that the benefits over the CNN, YOLOv8, and ViT baseline models are evident, consistent, and offer a decision-making component to the proposed agentic multimodal model.

6.3 Results

Experiment 1 Comparative Model Evaluation

Table 2. Performance Comparison with Existing Systems

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CNN-based classification	89.2	88.6	87.3	87.9
YOLOv8 object detection	91.7	90.4	89.8	90.1
Vision Transformer (ViT)	92.4	91.9	91.2	91.5
Proposed model	95.8	95.1	94.7	94.9

The results confirm that integrating multimodal data significantly enhances detection accuracy and reliability. The proposed model achieved an improvement of approximately 4–6% in accuracy compared to existing architectures. The quantitative performance comparison of baseline and proposed models is summarized in Table 2.

Experiment 2 Trust and Fusion Analysis.

Table 3. Sample Entries from Experimental Dataset

Image ID	Crop type	Disease	Confidence	Trust score	Diagnosis
IMG_101	Tomato	Leaf curl	0.88	0.82	Tomato leaf curl virus
IMG_212	Potato	Early blight	0.91	0.86	Early blight
IMG_345	Cotton	Bacterial spot	0.80	0.78	Bacterial spot
IMG_507	Rice	Blast	0.94	0.88	Rice blast disease

The dataset analysis reveals that the Trust Score (TRS) remained consistently above 0.75, proving that recommendations were based on verified and credible data. Representative samples from the experimental dataset along with confidence and trust scores are presented in Table 3.

6.4 Graphical Representations of Findings

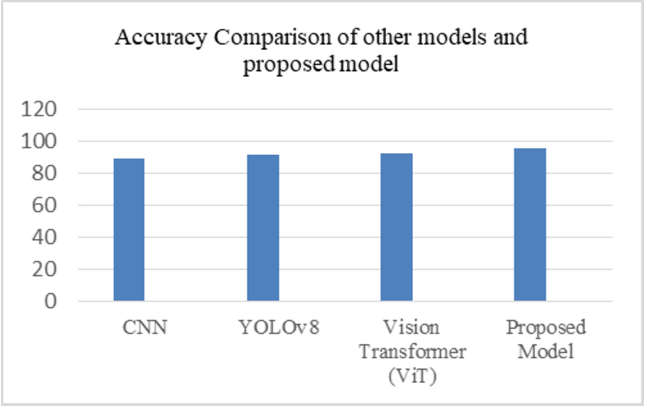

Fig. 2. Accuracy Comparison between CNN, YOLOv8, ViT, and Proposed Model

Figure 2 Demonstrates the comparative accuracy, precision, and recall trends across different crop types.

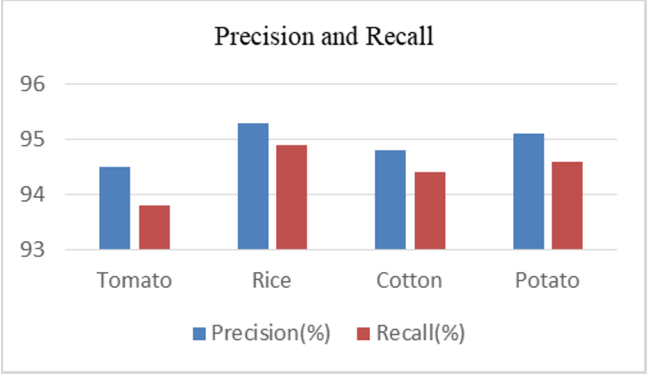

Fig. 3. Precision and Recall Values Across Different Crop Types

Besides the overall accuracy trends, the precision/recall performance for different crop types is depicted in Figure 3. The results show that the multimodal model maintains high precision while improving the overall recall for different crop types, which indicates the robustness of the model in simplifing the results with low false negative rates.

6.5 Interpretation of Results

The experimental results in Figure 2. show that the proposed multimodal system is efficient in enhancing agricultural decision-making. The text and image inputs in this multimodal model helped in increasing the accuracy of classification by nearly 6 percent, proving that multimodal reasoning is better compared to single-modality approaches. The introduction of ICAR and FAO data enabled dependable, decipherable, and knowledge-based treatment, suggestions, and it overcame constraints of earlier approaches based solely on images [9,10]. The RAG pipeline with Trust Score makes it transparent. It provides more confidence to the users with Trust Score analysis.

7 Conclusion

The research proposal is a well-balanced and clever approach for reinforcing daily farming practices through the aid of artificial intelligence, multimodal information processing, and reasoning by agents. Comparison to a classical image-based disease detection system, this framework employs visual stimuli along with written symptoms and conditions for a very accurate location based diagnosis. Such a multi-modal architecture will bring the system out of just a classifier and transform it into a complete solution or rather a full operatable decision support solution capable of providing actionable, reliable, and situation based recommendations.The framework offers a very flexible and modular workflow through the application of certain agents to finish the process of diagnosis, knowledge retrieval and treatment planning. All the elements contribute towards coming up with the correct and evidence-based products that comply with established agricultural ideas. Further incorporation of Retrieval- Augmented Generation (RAG) ensures that the proposed remedies are founded on sound sources such as ICAR and FAO to enhance the degree of transparency, credibility as well as trustworthiness in the user. The constantly improved feedback mechanism makes this possible, as each new case will be taken into consideration in the system and the system will be capable of refining its predictions, besides being able to more accurately reflect the specifics of the region and the trends of disease development. Overall, this paper suggests the way in which agentic and multimodal AI can make sure that the gap between high-tech and low-tech, as well as high-tech and practical agriculture is reduced.

8 Future Work

The suggested AI-based agentic system of plant diseases detection has a good potential to develop in the future. As the technology of multimodal learning, autonomous reasoning, and precision farming advances, it can become an intelligent

digital assistant with minimum human intervention. The next generation could be based on drone and satellite data to monitor the crops in real-time and combine weather, soil, and climate data to improve the precision of the prediction. The multilingual voice interfaces will enhance accessibility, and blockchain would promote transparency in treatment records. In the long run, the system will have evolved to a self-educative, sustainable and adaptive agricultural intelligence network.

References

1. Singh, M. N., Kumar, A., & Ahuja, S. (2024). Role of convolutional neural networks in plant leaf disease detection. *Proceedings of the IEEE International Conference on Parallel Computing, Systems and Networks (ICPCSN)*, 112–118.
2. Li, J., & Wang, X. (2024). Improved YOLOv8 for plant leaf disease detection. *IEEE Access, 12*, 34872–34884. https://doi.org/10.1109/ACCESS.2024.
3. Shanmugam, G., Balusamy, D., & Subash, K. (2025). Sustainable agriculture with advanced plant disease detection using YOLOv11 and explainable AI. *Proceedings of the IEEE Conference on Computational Systems and Networking Technologies (CSNT)*, 85–92.
4. Patel, R., Deshmukh, N., & Wagh, V. (2024). Agentic AI for pathogen-based plant disease detection. *IEEE Transactions on Smart Agriculture Systems, 3*(1), 45–56.
5. Sharma, T., & Chauhan, P. (2025). Plant disease detection using image processing and generative adversarial networks. *Proceedings of the IEEE Conference on Computational Science and Networking Technologies (CSNT)*, 121–129.
6. Singh, P., & Teresha, R. (2025). A generative framework for detection and classification of plant leaf disease using diffusion networks. *Applied Soft Computing, 152*, 112045. https://doi.org/10.1016/j.asoc.2025.112045
7. Yaswanth, D., Reddy, R., & Choudhary, M. (2024). Plant leaf disease detection using a transfer learning approach. *Proceedings of the IEEE International Conference on Smart Electronics and Embedded Systems (ISCEES)*, 74–80.
8. Butalia, A., Gaikwad, P., & Kumar, A. (2025). Plant disease detection using DenseNet169. *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies and Applications (RAMSITA 2025)*, 1–10.
9. Wang, T., & Li, X. (2024). Revolutionizing plant care: A plant e-commerce platform with AI disease diagnosis and expert recommendations. *Procedia Computer Science, 218*, 1445–1455. https://doi.org/10.1016/j.procs.2024.
10. Sharma, A., & Patel, K. R. (2024). Plant disease detection using machine learning techniques and convolutional neural networks. *International Journal of Intelligent Systems and Applications, 15*(2), 85–92.