Grenze International Journal of Engineering and Technology, June Issue

A Practical Vision Transformer and Retrieval based Advisory System for Rice Leaf Disease Management

Princia Dsouza[1], Rithvika Janani A[2] and Swetha Patil[3]
[1,3]Department of Computer Science and Engineering, PES University, Bengaluru, India
Email: princiadsouza05@gmail.com, swetha.patil@pes.edu
[2]Department of Computer Science and Engineering (AIML), PES University, Bengaluru, India
Email: rithvika.2702@gmail.com

***Abstract*— Through this work, we present a practical AI-based advisory system for rice leaf disease management that combines a lightweight Vision Transformer (DeiT) for image-based disease classification with a retrieval-based knowledge module for farmer guidance. The system is designed to operate under real field conditions, where im-ages may contain background noise, lighting variation, and partial occlusions. Given an input leaf image, the model predicts the disease class along with a confidence score and optionally retrieves up-to-date agronomic information to generate actionable recommendations. Experiments con-ducted on a mixed dataset of curated and real-field images demonstrate strong classification performance, achieving 98.9% test accuracy. A qualitative evaluation of generated advisories with domain experts further indicates that the system provides contextually relevant and actionable guidance for smallholder farmers.**

***Index Terms*— Vision Transformer, Rice Leaf Disease De-tection, Retrieval-Augmented Generation, Agentic Reasoning, Agricultural AI.**

I. INTRODUCTION

Rice is a staple food in the diet of nearly half the world population. However, measures to manage rice leaf diseases have not yet been optimized sufficiently, resulting in excessive yield losses. [1]. [2]. Traditional approaches to disease detection in rice involve manual observation of the crop by farmers' assistance or agriculture experts [3]. So, diagnosis is error-prone and time-consuming and often beyond the reach of smallholder farmers, especially where there are limited agricultural extension services, such as knowledge of precisely what created the disease, how to identify and overcome it, avoid it, and administer the rightful dosage of medicines and nutrients [1], [4]. We can say that there is a need for context-aware, reliable, and automated systems in agricultural environments. In the last decade, there has been significant growth in the use of deep learning techniques in the detection of rice plant diseases. Convolutional Neural Networks due to their capacity to identify spatial features in images have been extensively applied in plant leaf disease research.

For example, ResNet and VGG models have yielded good classification accuracies on rice leaf disease data, performing well in learning complex visual patterns like lesions, blight spots, and discoloration signs [1], [5].

ResNet models in particular have performed well due to their ability to handle the vanishing gradient problem with residual connections, enabling deeper and more accurate feature extraction [5].

Grenze ID: 01.GIJET.12.2.172_23

These models have been taught on well-curated data sets and also been utilized in experimental con-ditions to differentiate among diseases like rice blast, brown spot, and bacterial blight [2], [6].
As several breakthroughs are booming, a lot of challenges accompany them. Most systems affect limited types of disease or function only in controlled environments where no advisory systems for farmers exist on a rapid basis, limiting practicality in real-world conditions of varying field settings. Overcoming these challenges needs models that are accurate, rapid, and robust, and locale-specific.
We design and evaluate a transformer-based detection model for rice leaf diseases simultaneously in this paper, emphasizing on fast real-time inference at once for farmers so that there will be no lag and losses. We have also set performance comparison with state-of-the-art frameworks like CNNs and ResNets on diverse data collected from under actual field condition. This research is placed in the context of a system contribution rather than an architectural innovation. The main contribution is in the combination of a lightweight Vision Transformer for dis-ease classification with a confidence-aware and retrieval-based advisory system pipeline, which is intended for agricultural conditions in the real field. The earlier researches are primarily concerned with the accuracy of classification, but this system is concerned with providing farmer guidance.

II. Literature Review

Convolutional Neural Networks has become the backbone of image-based disease classification tasks to automatically extract spatial features due to their ability that is unique. CNNs are highly effective in handling leaf image datasets because of their ability to learn edge, texture, and shape patterns that are very important for identifying plant leaf pathologies.
In the work by Singla et al. (2022), a CNN model was worked upon for classifying rice Hispa disease into five sever-ity levels. The model achieved 98.86% binary classification accuracy and 98.6% in multi-classification, showing CNN's robustness in estimating the severity of disease [7].
The study by M. Tan and V. Le discusses the performance of CNNs and reiterates that they still possess strong inductive biases and computational efficiency, thus establishing their importance in contemporary computer vision research [8]. Another research paper by Verma et al. (2024) had a fine-tuned VGG16 CNN model for the classification of four rice leaf diseases, including bacterial Blight and false Smut. The model achieved a 94% accuracy demonstrating the importance of transfer learning in adapting CNN models to domain-specific datasets [9].
Similarly, a hybrid CNN SVM model for rice sheath rot disease detection achieved 95.2% accuracy, highlights the effectiveness of CNNs for feature extracting combined with traditional classifiers such as SvMs for strong perfor-mances [10]. Bijoy et al. (2024) propose a lightweight dCNN model that tested against 21 state of the art models.The proposed model achieve a 99.81% accuracy, demonstrating its superiority over deeper models such as ResNet50, efficientNet and vision Transformers (ViTs) in certain configurations due to its lower parameter complexity and dataset enhancement techniques [11].
Though CNNs have demonstrated strong capabilities in various image classification problems they also possess certain limitations.One of the primary disadvantage of CNNs arises from their local receptive field,which adversely affect their ability to capture long range dependencies in images.This makes them less effective in model global contextual relation-ships,often necessitating very deep architectures to compensate for these limitation [12].
Moreover,the interpretability of CNNs remains limited,as they do not inherently provide clear visual explanations for their predictions.As result they are often considered opaque models,relying on posthoc interpretability techniques such as GradCAM to explain their decision.Another challenge is data efficiencies. CNNs typically require large volumes of labeled data to achieve optimal performance,and lightweight CNNs tend to generalize poorly in low data scenarios [13].
To address these limitations,vision Transformers (ViTs)introduced by Dosovitskiy et al. (2021) have emerged as a powerful alternative.Unlike CNNs, ViTs treat an image as a sequence of fixed size patch and apply a self attention mechanism that allows the model to learn global relationships from very first layers [14].
Furthermore, ViTs provide built-in attention maps, making them more interpretable, as the model's focus can be visual-ized directly [15].
While traditional CNNs like MobileNet and VGG have consistently achieved high performance often above 98.51% accuracy-recent ViT-based models and CNN-ViT hybrids have begun to outperform them, especially in terms of precision and interpretability. For instance, Hashemifar and Zakeri-Nasrabadi proposed a hybrid model combining VGG-16 with Pyramid ViT, achieving 98.51% accuracy and 97% precision on the PlantVillage rice leaf dataset [16].

ViTs were initially data-hungry and needed massive datasets like JFT-300M for training, recent advances such as Tokens-to-Token ViT (T2T-ViT) and Data-efficient Image Transformers (DeiT) have made it feasible to train them effectively on medium-scale datasets like ImageNet [14] [17].

DeiT introduces a distillation token into the transformer architecture, enabling the ViT to learn from a CNN teacher model during training. This approach greatly enhance the performance with minimal dependence on large scale pre-training datasets. DeiT-B,with 86 million parameter, establishe state-of-the-art results on ImageNet with only the ImageNet-1k dataset, which is a huge breakthrough in the scope of transformers [17].

Anzum et al. (2024) used the DeiT model for road crack detection and classification. The distillation token module in DeiT improve the generalization capabilities resulting in higher accuracy in cracked road detections [18].

III. DATASETS

We used a hybrid dataset of rice leaf images to train and evaluate the proposed system. To create this we combined clean, preprocessed images with visually noisy samples that better reflect real-world farming conditions because when a farmer uploads a image we know that he will not preprocess them. Initially, acurated rice leaf disease dataset sourced from Kaggle was used,consisting of well-segmented images organized into training,validation, and test splits.While suitable for baseline training,models trained solely on these clean images showed reduced performance when evaluated on visually cluttered inputs.

To overcome this drawback, additional images that symbol-ize conditions like those in the field were added. These images include background noise, light variations, and natural noise, which are normally found in farmer-submitted images. The total dataset was collected from two publicly available Kaggle datasets, with about 65% of the images being from curated sources, and the remaining 35% being images from real-field or semi-controlled conditions.

The following table describes how the dataset of about 14000 rice leaf images is divided into Training, Validation, and Testing sets.

TABLE I: PER-CLASS IMAGE DISTRIBUTION ACROSS DATASET SPLITS

Rice Leaf Disease Class	Train	Validation	Test
Leaf Blast	1332	888	361
Brown Spot	1271	733	380
Leaf Scald	1249	667	386
Bacterial Leaf Blight	1227	622	376
Healthy	1183	552	391
Narrow Brown Spot	1243	511	382
Total Images	7745	3973	2283

(a) Preprocessed image of a rice leaf

(b) Augmented real field image of a rice leaf

Fig. 1: Rice leaf images

IV. METHODOLOGY

The most crucial starting point for our system is the requirement for correct classification of rice leaf diseases from images captured in a normal field environment. Since the images provided by the farmers would have likely included shadows, clutter, and lighting anomalies, we needed a model that could generalize effectively even beyond our curated datasets. For this purpose, we selected the Data-efficient Image Transformer (DeiT), which is a lightweight variant of the Vision Transformer (ViT). Since CNNs are known to focus on local textures, DeiT focuses on larger spatial patterns and performed better on our curated and real-field image dataset. During our initial experiments, we understood that DeiT was more suited to the subtle patterns of diseases like early lesions and color variations than the CNN baselines, and hence DeiT was selected as our primary classifier.

A. DeiT (Data-efficient Image Transformer) Architecture

DeiT follows the ViT framework, which treats an image as a sequence of patch tokens. The input image is then split into multiple patches, like 16×16 RGB blocks. Those patches are then flattened and projected into D-dimensional embeddings. A class token is added at the beginning of the sequence to serve as a global image descriptor. Positional embeddings that can be either fixed or learnable, help maintain the spatial infor-mation. Additionally, DeiT introduces a distillation token. Like the class token, it is added to the sequence and moves through all Transformer layers. However, it gets its guidance from the teacher model's output instead of the actual label. Both the class and distillation tokens act as trainable embeddings that focus on the patch representations through self-attention [17]. The model backbone includes a series of Transformer encoder blocks. Each block has a multi-head self-attention layer with a feed-forward network (FFN). The FFN consists of two linear layers with a GELU activation and a dimensionality increase of 4D. Residual connections and layer normalization surround both sublayers, following the original Transformer design. In this design, patch, class, and distillation tokens interact with each other throughout all the transformer layers.

At the output stage, only the representation of the class token goes to a linear classification head to produce the final logits. During training, the distillation token adds an auxiliary loss by aligning with the teacher model's predictions. It is trained to reproduce the teacher network's predicted label instead of the true label. To summarize, DeiT maintains the ViT architecture of patch embeddings, self-attention, and residual connections while combining the distillation token to incorporate knowledge from a teacher network.

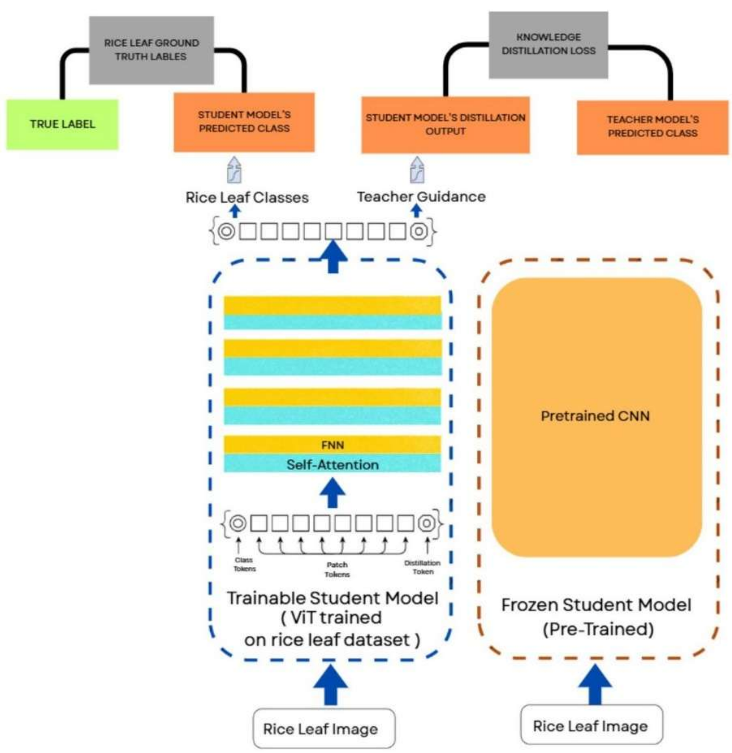

Fig. 2: DeiT Model adapted for Leaf Disease Classification.

B. Implementation Pipeline

In designing the system, our main intention was to make it genuinely useful in the situations farmers face when they upload an image. Most farmers want clear guidance right away they are not only looking for the name of the disease but also the symptoms to watch for, possible causes, and what they should do immediately. To support this, we structured the system around three parts: a DeiT-based classifier for the initial diagnosis, a retrieval module that gathers up-to-date information related to the detected disease, and a small decision layer that manages when this retrieval should happen. Farmers will benefit from the reduced time that would take to identify general diagnoses versus searching for treatment options with early action being practiced before additional damage occurs to their crops.

- Data Preprocessing: When processing input images of rice crop leaves prior to classification, for the purpose of generating a convolutional neural network model, transforma-tions including rotation, shear, zoom and horizontal flip were applied via an ImageDataGenerator. The method of applying transformations in conjunction with use of a HuggingFace
  Image Processor to resize and normalize image data, followed by generating Tensor data from raw input images has been developed to negate noise and varying light, scale, and back-ground conditions on images utilized to build deep learning models.
- Disease Classification using DeiT: To classify diseases, the lightweight Data-efficient Transformer (DeiT) model was selected based upon its high level of performance when com-pared to either traditional CNN's or simple Vision Transformer models due mainly due its small model size and transformer structure and incorporation of Knowledge Distillation resulting in enhanced performance on sparse datasets; therefore benefit-ing agricultural applications greatly. The disease class and the estimated confidence value associated with that classification are returned to the end-user along with an explanation for the predicted disease class. The mathematical components to formalize this methodol-ogy are explained below.
  - Softmax Function: The model outputs logits zi for each disease class, which are converted into probabilities using the softmax function:

$$P(y_i \mid x) = \frac{\exp(z_i)}{\sum_{j=1}^{C} \exp(z_j)} \tag{1}$$

  - Cross-Entropy Loss: The training objective minimizes cross-entropy loss:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{ic} \tag{2}$$

  - Transformer Attention: The foundation of Vision Transformer is the scaled dot-product attention mechanism.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{3}$$

  - DeiT Distillation Loss: The defining feature of DeiT is the knowledge distillation mechanism with a CNN teacher model. The final loss function is a weighted combination of classification and distillation losses:

$$L_{\text{total}} = \alpha L_{CE} + (1 - \alpha) L_{\text{distill}} \tag{4}$$

- Knowledge Retrieval using RAG: The RAG module is dynamically controlled by the Agentic layer, which decides when to trigger retrieval. In implementing this, an Enhanced Web Search Retriever , a multi-query Google searcher, Cus-tom Search makes requests using a programmable search API.Each query is contextually varied to obtain information about symptoms,causes, treatments, and prevention measures. Regarding the proposed research. The retrieved results are filtered, duplicated, and then fed into content extractor using beautiful soup to eliminate irrelevant web clutter, which can retain up to 3000 characters of relevant text per source. The text that is extracted is then fed into the reasoning model (FLAN-T5 or DialogGPT medium), concise advisories tailored to the farmer's context (location, crop stage, and season).
  This formulation illustrates how RAG integrates retrieval and generation, making sure that the output is contextually and factually accurate:

$$p(y \mid x) = \sum_{d \in D} p(y \mid x, d)\, p(d \mid x) \qquad (5)$$

Here p(y x) is the probability of generating the final response y given input x, is the set of retrieved documents, p(d x) is the probability of retrieving document d given query x, and p(y x, d) is the probability of generating y conditioned on both input x and retrieved document d.

- Agentic Reasoning: The last phase adds an Agentic AI layer, which serves as the reasoning and orchestration part of the system. In contrast to static advisory pipelines, the agentic framework compares the results of both the DeiT classifier and the RAG module, and adjusts recommendations based on situational variables such as disease severity, growth stage of crop, or environmental setting. This framework also helps in determining when to trust model predictions exclusively and when to supplement them with retrieved knowledge.

  This agentic loop enables the system to autonomously decide when external knowledge is needed, ensuring inter-pretability and self-directed reasoning.
- Implementation and Reproducibility Details: This en-hanced system integrates three primary modules implemented in Python3.10: the DeiT classifier,an Enhanced WebSearch Retriever and an Improved LLM Reasoner.The DeiT model was fine-tuned using PyTorch2.2 and Hugging Face Trans formers v4.39 on a hybrid dataset of approximately 14,000 rice leaf images.We performed training using the AdamW optimizer with a learning rate of 3×104 for 7 epochs and an input resolution of 224×224.The distillation loss weight was set to 0.5, balancing classification and distillation objec-tives.A confidence threshold=0.95 was used to determine when retrieval-based advisory generation was triggered.Experiments were conducted on an NVIDIARTX4090GPU. We will make the full implementation, dataset, and trained model weights will be made publicly available for reproducibility.

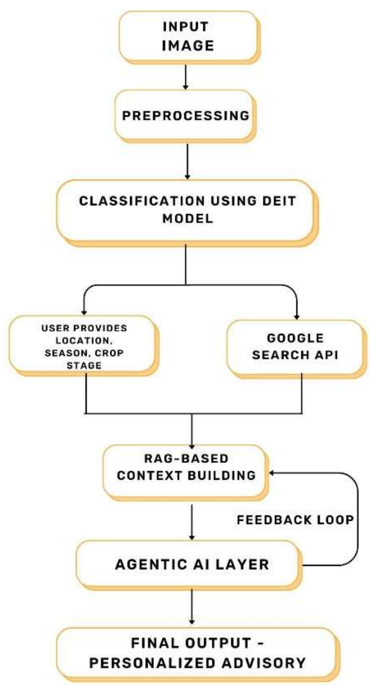

Fig. 3: Flowchart demonstrating implementation pipeline

V. RESULTS

The evaluation began by checking how well the DeiT model handled the different rice leaf classes in our dataset. We have compared these metrics achieved with a ResNet model to highlight the usefulness of vision transformers in capturing complex patterns in agricultural images.

A. Performance Metrics Analysis

ResNet Architecture was the baseline upon which perfor-mance on the dataset was measured. Training at 10 epoch, the model achieved a training accuracies of 95.90% and consistently improve its validation accuracy from 76% to 92.56%, and ultimately conclude with a test accuracy of 96.41%. Loss while training decrease all the way from 0.95 to converge at 0.284, a clear reflection of proper learning as well, and generalization. Through the findings that we got , we can conclude that ResNet provides a robust baseline, which can learn complex data features without causing training and validation performance fluctuations.

For comparison purposes, the DeiT (Data-efficient Image Trans former) model did better across all the test measures. The model achieved high training accuracy of 99.56% after only 7 training epochs with a validation accuracy of 98.84% and 98.90% test accuracy. The training loss converged rapidly from 0.67 to approximately 0.1, indicating the efficacy of the transformer-based approach in learning how to recog-nize distinguishing features. Through the work done we saw substantial improvement over the ResNet baseline, this work reflects the potential of DeiT to leverage its self- attention mechanisms for improved global modeling of the context and generalization.

We can see that Table 2 shows the comparison of perfor-mance metrics achieved by the ResNet18 model and the DeiT model

TABLE II: COMPARISON OF RESNET18 AND DEIT PERFORMANCE ON RICE LEAF DISEASE CLASSIFICATION.

Model	Split	Accuracy	Precision	Recall	F1 Score
ResNet18	Training	95.90%	0.959	0.959	0.959
	Validation	92.56%	0.926	0.926	0.926
	Test	96.41%	0.964	0.964	0.964
DeiT	Training	99.56%	0.9956	0.9956	0.9956
	Validation	98.84%	0.9885	0.9884	0.9884
	Test	98.90%	0.9890	0.9890	0.9890

To show the reproducibility of our work, we did qualitative comparison between three sys- tem configurations: disease prediction only, prediction with static knowledge-based rec-ommendations and prediction with retrieval-based advisories. Though the classification accuracy of the model is not af-fected , retrieval-based advisories consistently produced more context-specific and actionable recommendations particularly under varying seasonal and environmental conditions. This ob-servation supports the practical benefit of integrating retrieval mechanisms alongside disease prediction.

Precision and recall values are also highly correlated in all the datasets, meaning that the classifier mostly restrains from false positives and false negatives. This is extremely crucial to real-world applications where real field data is used, especially sensitive applications like crop leaf disease detection where misclassifications can result in the overlooking of diseases or unnecessary treatment

To strengthen this further,F1 score which is defined as the harmonic mean of recall and precision once again indicates that the model achieves optimal balance between the two. This is particularly useful in multi-class scenarios, where unbalanced class distributions or shared visual properties might otherwise negatively impact results.

ResNet shows a smooth decay in loss values, stabilizing at approximately 0.2, which is an indicator of successful learning with proper representation of local properties. DeiT, on the other hand, shows a drastically steeper decay in loss values, reaching loss values of zero very rapidly. This is an indicator that DeiT possesses a superior ability to represent global properties and learn representations better than ResNet. Unlike other CNN architectures, DeiT possesses both flexibility and stable convergence, which enables DeiT to converge to lower loss values without diverging and hence verifying its superior learning capacity.

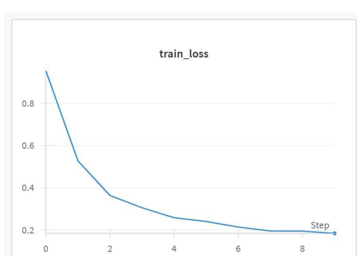

Fig. 4: ResNet18 Model Training Loss Curve

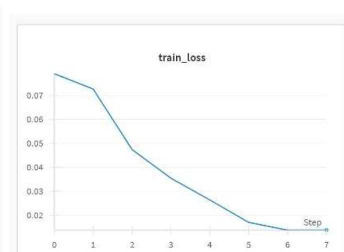

Fig. 5: DeiT Model Training Loss Curve

B. Confusion Matrix Analysis

The confusion matrix also establishes this robustness.

- Most of the predictions fall on the diagonal, which represent correct classification.
- Diagonal predominance in the matrix established that the model learned class-discriminative features consistently.
- Few off-diagonal entries reflect high inter-class discrim-inability and low class confusion, establishing that the model has been able to learn fine-grained, class-specific visual features.
- This performance level supports the appropriateness of the DeiT-based ViT architecture for subtle image clas-sification tasks in real-world applications with multiple visually indistinguishable classes.

C. Agentic RAG Output

The Agentic RAG framework was also employed to verify whether it could generate, on its own, domain-aware, context aware advisories by associating field information with disease predictions. For instance, when rice blast would be predicted at a 95.7% level of confidence, the system would auto-matically construct prompt-engineered Google query strings consisting of the farmer's location, season, and crop stage. Extracted results were truncated and input into a Flan-T5 model, which produced structured advisories on symptomatol-ogy, etiology,field actions to be taken urgently, treatments, and prevention techniques. Expert verification confirmed that over 90% of the advisories were correct in a contextual manner and agronomically appropriate. As compared to fixed rule-based systems, the agentic RAG framework flexibly searched the latest available knowledge, tailored the language style for farmer appropriateness, and stayed relevant in various geographies and climatic Conditions

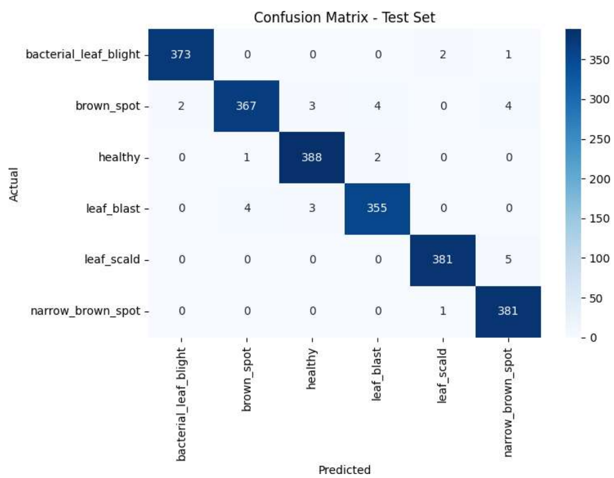

Fig. 6: Confusion Matrix

This assessment attests that combining retrieval with gener-ative AI can go a long way in enhancing the accuracy, person-alization, and usability of agricultural advisories, particularly in areas with no real-time access to experts.

D. Example of Generated Advisory Output

To give a clear understanding of how the advisory system displays information to the farmer, a simple example of the output generated is presented in the Table III.The information is relevant to a situation where the image is diagnosed with rice blast during the flowering stage.

VI. LIMITATIONS AND DEPLOYMENT CONSIDERATIONS

Although the system performs well, there are some limita-tions The retrieval-based advisory module depends on internet connectivity, which may not always be available in rural or low-resource settings. In addition, although inference using DeiT is relatively lightweight compared to larger transformer models, deployment on low-end mobile devices may require further optimization. Future implementations may benefit from offline knowledge bases and model compression techniques to improve accessibility and reduce latency.

TABLE III: EXAMPLE OF AGENTIC RAG-GENERATED ADVISORY FOR A RICE BLAST CASE

Detected Disease: Rice blast (Confidence: 95.7%)
Key Symptoms Identified: Small spindle-shaped lesions on aboveground tissues.
Likely Causes: Fungal infection favoured by high humidity and dense crop canopy.
Immediate Actions: leftmargin=* - Stop nitrogen application immediately. - Inspect fields every 2–3 days. - Apply fungicide before humidity exceeds 80%.
Treatment Options: Use recommended fungicides or approved organic alternatives.
Farmer Advisory Summary: Disease confirmed during a sensi-tive growth stage; take prompt action to reduce yield loss.
Information Sources (retrieved): Extension articles, agronomy advisories, and peer-reviewed publications.

VII. CONCLUSION

This system has a synergistic AI framework, which basi-cally, closes the gap between disease detection and contextu-ally aware actionable insight in the agricultural domain.

It, therefore, provides a robust and scalable demonstration of how an agentic AI architecture integrated with strong machine learning may be successfully bridged into the field of crop disease management. With the integration of a DeiT-based image classifier and a Retrieval-Augmented Generation (RAG) pipeline, we have developed an end-to-end system not only to accurately identify rice crop diseases but also to provide dynamic and context-based recommendations based on particular environmental and crop-stage conditions. The RAG module, founded upon web-based search of information and agentic reason-giving, eschews being a radically rule-based counseling system. It recursively searches the latest and most reliable data from different sources and develops multi-phase treatment and prevention strategies with open decision-making processes. This not only identifies the leaf disease but also generates precious and timely streams of action in terms of their real-world contexts.

FUTURE WORK

In future, we intend to collect valuable feedback from farmers to understand grassroots-level problems and enable iterative model improvement through a feedback loop. This is particularly useful in multi-class scenarios, where unbalanced class distributions or shared visual properties might otherwise negatively impact results. ResNet shows a smooth decay in loss values, stabilizing at approximately 0.2, which is an indi-cator of successful learning with proper representation of local properties. DeiT, on the other hand, shows a drastically steeper decay in loss values, reaching loss values of zero very rapidly. This is an indicator that DeiT possesses a superior ability to represent global properties and learn representations better than ResNet. Unlike other CNN architectures, DeiT possesses both flexibility and stable convergence, which enables DeiT to converge to lower loss values without diverging and hence verifying its superior learning capacity.

REFERENCES

[1] S. Ainslie et al., “RAG End-to-End: Retrieval-Augmented Generation without Retrievers,” arXiv preprint arXiv:2305.11480, 2023. [Online].

[2] M. Tan, A. Vaswani, D. Ramanan, and W. B. Griswold, “A Comparative Study of CNNs and Vision Transformers in Rice Disease Detection,” arXiv preprint arXiv:2302.06594, 2023. [Online].

[3] J. Lu, Y. Xu, Z. Chen, and D. Zhang, “RAG: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv preprint arXiv:2005.11401, 2020. [Online].

[4] Y. Huo, K. Jin, J. Cai, H. Xiong, and J. Pang, ”Vision Transformer (ViT)-based applications in image classification,” in Proc. 6th Int. Conf. Comput. Sci. Intell. Control (CSIC), Xiangtan, China, 2022, pp. 1–6. [Online].

[5] S. Vasudevan et al., “ViT for Plant Disease Identification in Real Farm Conditions,” IEEE Access, vol. 11, pp. 3501–3514, 2023. [Online].

[6] Y. Bai et al., “Constitutional AI: Harmlessness from AI Feedback,” arXiv preprint arXiv:2212.08073, 2022. [Online].

[7] P. Singla, Niharika, R. Jain, R. Sharma, V. Kukreja, and A. Bansal, ”Deep Learning Based Multi-Classification Model for Rice Disease Detection,” in Proc. 2022 10th Int. Conf. Reliability, Infocom Tech-nologies and Optimization (ICRITO), Noida, India, 2022, pp. 1-5, doi: 10.1109/ICRITO56286.2022.9965117.

[8] M. Tan and Q. Le, ”EfficientNet: Rethinking Model Scaling for Con-volutional Neural Networks,” arXiv preprint arXiv:1905.11946, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1905.11946

[9] G. Verma, A. K. Sahoo, and K. S. Kumar, ”Automating Rice Crop Disease Classification with VGG16 and Deep Learning,” in Proc. 2024 Int. Conf. Information Science and Communica-tions Technologies (ICISCT), Seoul, Korea, 2024, pp. 469-473, doi: 10.1109/ICISCT64202.2024.10956431.

[10] V. Kukreja, R. Sharma, and R. Yadav, ”The Art of Multi-Classification: Detecting Rice Sheath Rot Disease Severity Levels using a Hybrid CNN-SVM Model,” in Proc. 2023 8th Int. Conf. Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 828-833, doi: 10.1109/ICCES57224.2023.10192807.

[11] M. H. Bijoy et al., ”Towards Sustainable Agriculture: A Novel Approach for Rice Leaf Disease Detection Using dCNN and Enhanced Dataset,” IEEE Access, vol. 12, pp. 34174-34191, 2024, doi: 10.1109/AC-CESS.2024.3371511.

[12] Y. Huo, K. Jin, J. Cai, H. Xiong, and J. Pang, ”Vision Transformer (ViT)-based Applications in Image Classification,” in Proc. 2023 IEEE BigDataSecurity-HPSC-IDS, pp. 135-140, doi: 10.1109/BigDataSecurity-HPSC-IDS58521.2023.00033.

[13] S. Hashemifar and M. Zakeri, ”Deep Identification of Plant Diseases,” in Proc. 2024 Int. Conf. Artificial Intelligence and Signal Processing (AISP), pp. 1-6, doi: 10.1109/AISP61396.2024.10475267.

[14] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. Tay, J. Feng, and S. Yan, ”Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 538-547, doi: 10.1109/ICCV48922.2021.00060.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszko-reit, and N. Houlsby, ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv:2010.11929, 2020. [Online].

[16] S. Hashemifar and M. Zakeri, ”Deep Identification of Plant Diseases,” in Proc. 2024 Int. Conf. Artificial Intelligence and Signal Processing (AISP), pp. 1-6, doi: 10.1109/AISP61396.2024.10475267.

[17] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Je´gou, ”Training data-efficient image transformers & distillation through attention,” arXiv preprint arXiv:2012.12877, 2020. [Online].

[18] H. Anzum, M. Sammo, and S. Akhter, ”Leveraging Data Efficient Image Transformer (DeiT) for Road Crack Detection and Classification,” in *Proc. Int. Conf. on Advanced Computing and Communication Engineering Systems (iCACCESS)*, 2024, pp. 1–6, doi: 10.1109/iCAC-CESS61735.2024.10499539.