International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

DIEC-ViT-Driven Tobacco Leaf Disease Diagnosis and Product Retrieval Using AI

1st Dr.G.V.S.N.R.V Prasad
Professor, Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Andhra Pradesh, India
gutta.prasad1@gmail.com

2nd B.Haritha Sai
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Andhra Pradesh, India
barmaharithasai@gmail.com

3rd A.Yaswanth Kiran
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Andhra Pradesh, India
yaswanthjk580@gmail.com

4th Ch.Mounika
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Andhra Pradesh, India
chandumounika403@gmail.com

5th G.Kartheek
Department of Computer Science and Engineering
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Andhra Pradesh, India
gosalakartheek0704@gmail.com

***Abstract*—Tobacco crop productivity is significantly affected by leaf diseases, which often remain undetected at early stages due to strong visual similarity among symptoms. Automated disease identification can assist farmers in timely diagnosis and effective disease management. This paper proposes an end-to-end tobacco leaf disease detection and advisory framework that integrates deep learning–based visual analysis with knowledge-driven information retrieval. Discriminative Information Enhanced Contrastive Vision Transformer (DIEC-ViT) is employed for disease classification using a tobacco-specific leaf image dataset. The proposed model achieves a test accuracy of 88.61%, outperforming baseline Convolutional Neural Network (CNN), ResNet, and Recurrent Neural Network (RNN) models in terms of accuracy, convergence stability, and generalization performance. To enhance practical usability, a structured tobacco disease knowledge base is semantically encoded using sentence embedding algorithms and indexed with FAISS for efficient similarity-based retrieval. A retrieval-augmented generation (RAG) mechanism is further incorporated to ensure that advisory responses remain factual, reliable, and explainable. Experimental results demonstrate that the integration of transformer-based visual modeling with semantic retrieval significantly improves both disease recognition and decision-support capability. The proposed framework offers a robust and deployable solution for intelligent tobacco disease management in real-world agricultural settings.**

***Index Terms*—Vision Transformer, Tobacco Leaf Disease Detection, Plant Disease Classification, Retrieval-Augmented Generation, FAISS, Precision Agriculture**

I. Introduction

Tobacco is a highly valuable commercial crop whose yield and quality are significantly affected by foliar diseases. Diseases such as Brown Spot, Frog Eye, Mosaic, Potato Virus Y, Alternaria Alternata, and Cercospora Nicotianae often show similar early symptoms, making manual diagnosis difficult. Incorrect or delayed identification can lead to improper treatment, excessive pesticide use, and substantial yield loss.

Recent developments in computer vision and deep learning have enabled automated plant disease detection using leaf images. While Convolutional Neural Networks (CNNs) are widely used, they often struggle to capture global spatial relationships. Vision Transformers (ViTs), with their self-attention mechanism, overcome this limitation by modeling long-range dependencies and have shown strong performance on plant disease datasets.

However, most existing approaches focus only on disease classification and lack practical advisory support. Farmers require actionable insights such as symptoms, causes, and treatment recommendations. Additionally, many models are trained on generic datasets and do not address tobacco-specific disease characteristics.

To address these limitations, this work proposes an intelligent tobacco leaf disease detection and advisory system. The system integrates a custom-trained Vision Transformer with a semantic retrieval framework. A tobacco-specific dataset is used to train the model, while preprocessing techniques ensure robust performance.

A structured knowledge base is developed containing disease descriptions, symptoms, causes, and treatment methods. Sentence embeddings and FAISS-based similarity search are used to enable efficient semantic retrieval. A retrieval-augmented generation (RAG) approach ensures that responses are accurate and grounded in the knowledge base. A tool-based agent coordinates image-based disease detection and

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

text-based query handling.

The system is deployed using a Flask-based web interface, allowing users to upload images or ask queries interactively. By combining classification with explainable advisory support, the proposed system provides a practical decision-support tool for tobacco disease management and can be extended to other crops.

II. Related Work

Automated plant disease detection using image-based analysis has received significant attention due to its ability to improve crop health monitoring and reduce yield losses. Early approaches employed traditional machine learning techniques with handcrafted features such as color histograms, texture descriptors, and shape features [1]. Although effective in controlled settings, these methods lacked robustness to illumination variations, background clutter, and visually similar disease symptoms. The introduction of deep learning, particularly Convolutional Neural Networks (CNNs), marked a major advancement in plant disease classification. Mohanty et al. demonstrated that deep CNN architectures significantly outperform traditional methods on leaf disease datasets [2]. Several studies extended CNN-based models to tobacco crops, reporting improved classification accuracy using field-acquired images [3], [4]. However, CNNs mainly learn local spatial features and often fail to capture long-range dependencies when disease symptoms are irregularly distributed across leaf surfaces. To overcome these limitations, deeper architectures such as Residual Networks (ResNet) were introduced to enhance feature reuse and generalization [5], [6]. More recently, Vision Transformers (ViTs) have emerged as a powerful alternative by modeling global contextual relationships through self-attention mechanisms [7]. ViT-based models have demonstrated superior performance over CNNs in plant disease classification, particularly for visually subtle and spatially dispersed symptoms [8], [9]. Nevertheless, most existing ViT-based approaches rely on generic plant datasets and lack crop-specific adaptation for real-world agricultural deployment.

In parallel, agricultural decision-support systems have increasingly integrated natural language processing techniques to complement visual disease detection. Sentence embedding methods enable the conversion of unstructured disease descriptions into dense semantic representations while preserving contextual meaning [10]. Efficient similarity search libraries such as FAISS support fast and scalable retrieval of relevant disease information from large knowledge bases [11]. Furthermore, retrieval-augmented generation (RAG) frameworks have been proposed to ground language model outputs in retrieved factual content, thereby improving the reliability and explainability of generated recommendations [12].

Beyond agriculture, optimization-enhanced machine learning and hybrid deep learning frameworks have demonstrated improved robustness and predictive performance in medical diagnostic systems. Cuckoo search and bat search optimization–based feature selection techniques have been successfully applied to complex classification problems, highlighting the benefits of hybrid optimization strategies [13], [14]. Additionally, hybrid CNN and machine learning models using deep visual features have shown strong performance in real-world identification tasks [15]. Feature representation and similarity-based retrieval methods [16], along with optimization-enabled deep learning frameworks for scalable decision-support systems, further support the integration of advanced learning, retrieval, and optimization techniques [17].

Despite these advances, most tobacco disease detection studies remain limited to disease classification and do not extend to advisory or recommendation systems. This research addresses this gap by proposing a tobacco-specific Vision Transformer–based disease detection framework integrated with sentence embedding–based semantic retrieval and retrieval-augmented generation, enabling accurate diagnosis alongside reliable, knowledge-driven decision support.

III. Proposed Methodology

This section presents the detailed methodology of the proposed tobacco leaf disease detection and advisory system. The overall framework consists of four major stages: image preprocessing, Vision Transformer–based disease classification, semantic knowledge representation and retrieval, and intelligent response generation. Each stage is designed to operate independently while contributing to a unified end-to-end decision support system.

The proposed system follows a modular architecture consisting of four primary components:

(i) Image preprocessing and disease detection
(ii) Disease knowledge representation and semantic embedding
(iii) Similarity-based information retrieval and response generation
(iv) Intelligent agent orchestration and user interaction.

User inputs in the form of leaf images or textual queries are processed through the appropriate computational pipeline. Image inputs trigger the Vision Transformer–based disease classification module, while text-based queries invoke semantic retrieval from the disease knowledge base. The system ensures seamless integration between vision-based detection and language-based advisory generation.

A. Dataset overview

The proposed system employs two complementary datasets to support both visual disease detection and knowledge-driven advisory generation. The primary dataset consists of tobacco leaf images obtained from the Roboflow Tobacco Leaf Disease Classification dataset. This dataset includes 2,540 original images, which were expanded to 15,240 images through extensive data augmentation techniques such as rotation, flipping, brightness variation, contrast adjustment, and zooming to improve model generalization and mitigate class imbalance. All images were standardized to a resolution of 640 × 640 pixels and categorized into ten classes: Alternaria alternata, Brown spot, Cercospora nicotianae, Frog eye, Healthy, Mosaic, No cercospora nicotianae, Potato virus Y, Target spot, and Weather fleck. The dataset captures diverse disease symptoms

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

under varying illumination, background, and leaf orientation conditions, enabling robust training and evaluation of the Vision Transformer–based classification model.

In addition to the image dataset, a structured tobacco disease knowledge base was developed to support recommendation and advisory generation. This secondary dataset contains curated textual information for each disease class, including disease descriptions, symptom characteristics, occurrence conditions, preventive measures, and natural treatment recommendations. The knowledge dataset is processed using sentence embedding techniques and indexed for efficient semantic retrieval, allowing the system to provide explainable and context-aware guidance following disease detection. The combined use of a tobacco-specific image dataset and a domain-driven knowledge base enables the proposed framework to perform accurate disease classification while delivering actionable decision support, thereby extending beyond conventional image-based plant disease detection approaches.

1) Preprocessing: All input tobacco leaf images undergo a standardized preprocessing pipeline prior to model training and inference. Initially, images are resized to a fixed resolution of 224×224 pixels to satisfy the input requirements of the Vision Transformer architecture. Pixel intensity values are normalized using channel-wise mean and standard deviation to reduce illumination variations and stabilize model optimization. The preprocessed images are then converted into tensor representations suitable for deep learning frameworks.

Formally, given an input image $I \in \mathbb{R}^{H \times W \times C}$, the preprocessing operation is defined as:

$$I' = \mathrm{P}(I) = \mathrm{Normalize}\ \mathrm{Resize}(I, 224 \times 224) \quad (1)$$

Prior to training, offline data augmentation is applied to the original dataset to expand its size and diversity. Geometric transformations include random rotation within $\pm 25^{\circ}$, horizontal and vertical flipping, and shift–scale–rotate operations to simulate variations in leaf orientation, position, and scale. Photometric augmentations such as random brightness and contrast adjustment are used to model illumination changes. Additionally, Gaussian noise injection and blur augmentation are employed to improve robustness against sensor noise and focus variations. The augmentation process can be expressed as:

$$\tilde{I} = \mathrm{A}(I') = \mathrm{T}(I'; \vartheta) + \epsilon \quad (2)$$

where $\mathrm{T}(\cdot)$ represents the applied spatial and photometric transformations parameterized by ϑ, and ϵ denotes additive noise. This offline augmentation strategy is performed before model training to construct an enriched dataset, enabling the Vision Transformer to learn invariant and discriminative features and thereby improving generalization performance in tobacco leaf disease classification.

B. Vision Transformer–Based Disease Detection

In the proposed system, tobacco leaf disease classification is performed using a Vision Transformer (ViT) architecture that models an image as a sequence of visual tokens. Each

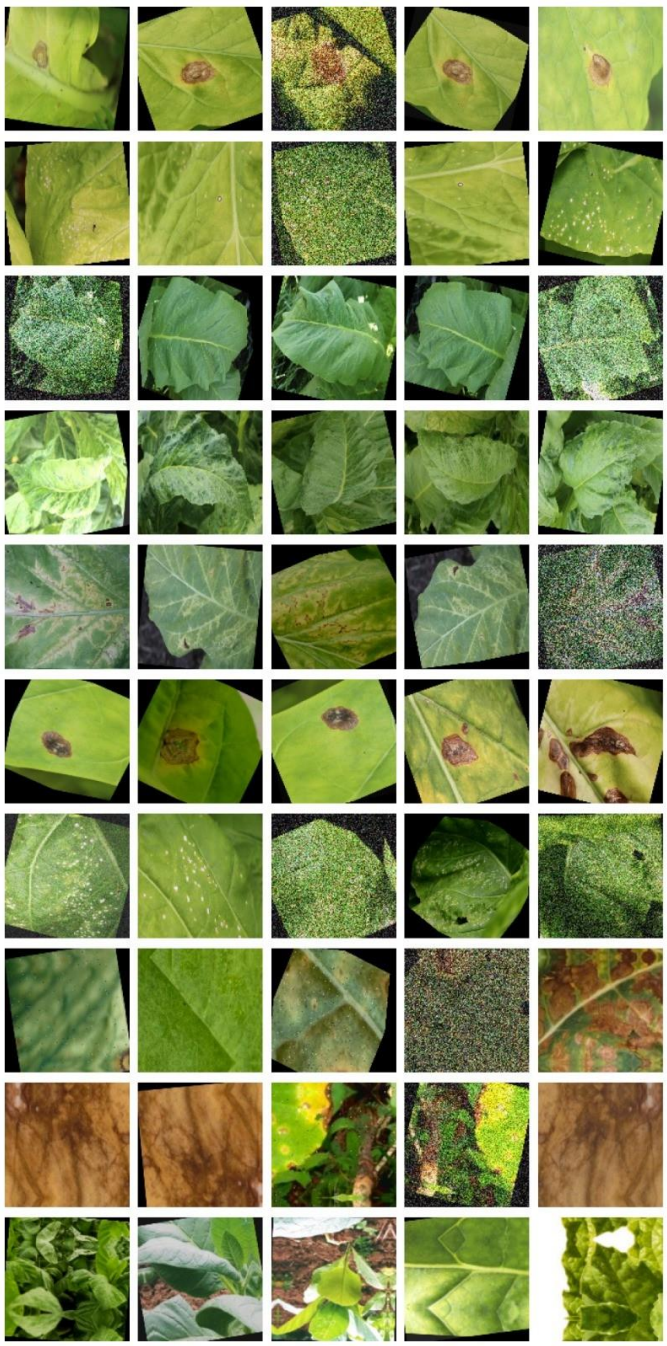

Fig. 1. Representative samples from the tobacco leaf disease dataset

input image $I \in \mathbb{R}^{H \times W \times C}$, after undergoing preprocessing and normalization, is resized to a fixed spatial resolution of 224×224 pixels. The resized image is then divided into a set of non-overlapping square patches of size $P \times P$, where $P = 16$. The total number of patches N generated from an image is given by

$$N = \frac{H \times W}{P^2} \quad (3)$$

which results in $N = 196$ patches for the adopted configuration. Each patch captures local visual information from

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

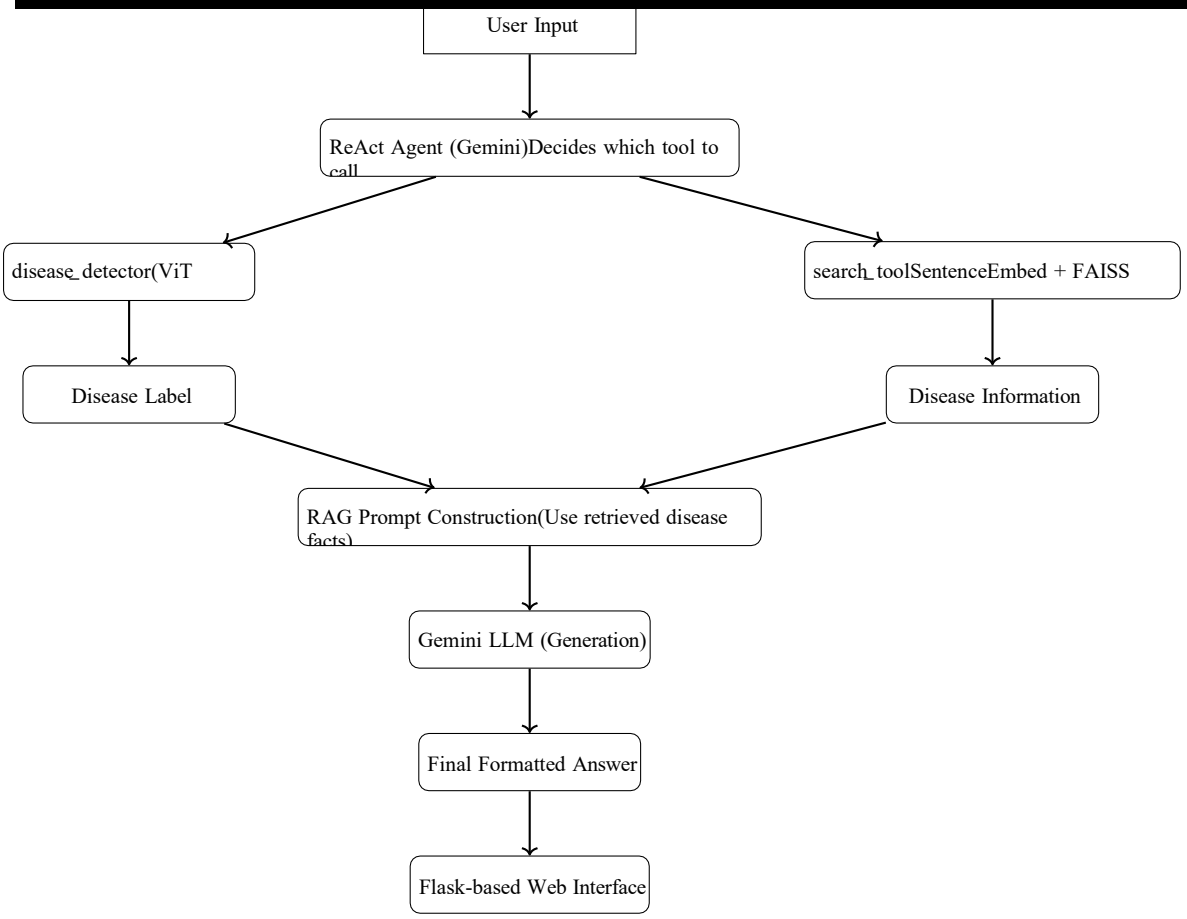

Fig. 2. Flowchart of the proposed ViT-based disease detection and retrieval-augmented advisory framework

a distinct region of the tobacco leaf while enabling global reasoning in subsequent stages. Each image patch is flattened into a one-dimensional vector $x_i \in R^{P^2C}$ and projected into a latent embedding space of dimension D through a trainable linear transformation, defined as

$$z_i = x_i E \tag{4}$$

where $E \in R^{(P^2C)\times D}$ denotes the learnable patch embedding matrix. This embedding process maps all patches to a uniform feature space, allowing them to be processed as a sequential input. Since transformer models lack inherent spatial awareness, positional information is explicitly incorporated by adding learnable positional embeddings. In addition, a learnable classification token $z_{cls} \in R^D$ is prepended to the patch sequence. The resulting input sequence to the transformer encoder is expressed as

$$Z_0 = [z_{cls}; z_1; z_2; \ldots; z_N] + E_{pos} \tag{5}$$

where $E_{pos} \in R^{(N+1)\times D}$ represents the positional embedding matrix. This formulation preserves spatial relationships while enabling global attention across the image. The embedded sequence is processed through a stack of transformer encoder layers, each consisting of a multi-head self-attention (MHSA) mechanism followed by a position-wise feed-forward network. For an encoder input $Z \in R^{(N+1)\times D}$, the query, key, and value matrices are computed as

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \tag{6}$$

where $W_Q, W_K, W_V \in R^{D\times d_k}$ are learnable projection matrices. The self-attention operation is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\frac{QK^T}{\sqrt{d_k}} V \tag{7}$$

This mechanism enables each patch to attend to all others, allowing the model to capture long-range spatial dependencies and contextual relationships between distant disease regions. Multiple attention heads operate in parallel to learn diverse feature interactions. Following the self-attention block, a position-wise feed-forward network (FFN) enhances non-linear feature learning and is defined as

$$\text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \tag{8}$$

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

Residual connections and layer normalization are applied around both the self-attention and feed-forward modules to stabilize training. The output of the l-th encoder layer is computed as

$$Z_l = \mathrm{FFN}(\mathrm{MHSA}(\mathrm{LN}(Z_{l-1}))) + Z_{l-1} \quad (9)$$

After passing through all transformer encoder layers, the final representation corresponding to the classification token z_{cls} is extracted. This vector aggregates global contextual information from all image patches. The classification head consists of a fully connected layer followed by a softmax activation:

$$\hat{y} = \mathrm{softmax}(W_c z_{\mathrm{cls}} + b_c) \quad (10)$$

where W_c and b_c are learnable parameters, and $\hat{y}$ denotes the probability distribution over tobacco disease classes.

C. Semantic Representation of Disease Information

In the proposed framework, disease-related knowledge is represented in a semantic vector space to enable meaning-based retrieval rather than keyword-based matching. The textual knowledge base consists of structured entries describing tobacco diseases, including symptoms, occurrence conditions, preventive measures, and natural treatment recommendations. These textual entries are unstructured in nature and therefore require transformation into a numerical form that preserves semantic meaning for computational processing.

To achieve this, each disease-related text s is encoded into a dense, fixed-dimensional vector representation using a sentence embedding algorithm. The embedding function $\phi(\cdot)$ maps a text sequence to a vector $\phi(s) \in \mathbb{R}^d$, where d denotes the embedding dimension. This transformation is learned such that semantically similar texts are located closer to each other in the embedding space, while unrelated texts are positioned farther apart. As a result, disease descriptions sharing similar symptoms or treatment strategies exhibit high similarity even if they differ in vocabulary or phrasing.

In the implemented system, all disease descriptions and advisory texts in the knowledge base are preprocessed and encoded offline to generate their corresponding embedding vectors. These embeddings form a compact semantic index that captures the contextual relationships between different tobacco diseases and their associated management practices. When a disease is detected by the Vision Transformer model, a query text corresponding to the predicted disease name or symptom description is encoded using the same embedding function, ensuring consistency within the semantic space.

The relevance between a query text s_1 and a candidate knowledge entry s_2 is computed using cosine similarity, defined as:

$$\mathrm{Sim}(s_1, s_2) = \frac{\phi(s_1) \cdot \phi(s_2)}{\|\phi(s_1)\| \, \|\phi(s_2)\|} \quad (11)$$

Cosine similarity measures the angular distance between two embedding vectors and is invariant to vector magnitude, making it well suited for comparing high-dimensional semantic representations. Higher similarity values indicate greater semantic relevance between the query and the retrieved disease information.

In the proposed system, similarity scores are used to rank all disease knowledge entries, and the top-matching entries are selected as relevant advisory content. This retrieval process ensures that the generated recommendations are grounded in semantically related, crop-specific information rather than generic or unrelated text. By leveraging dense semantic representations and cosine similarity-based retrieval, the system achieves accurate, explainable, and context-aware access to tobacco disease knowledge, forming a critical component of the overall retrieval-augmented advisory framework.

D. FAISS-Based Similarity Search

To enable efficient and meaning-based retrieval of tobacco disease information, all textual entries in the disease knowledge base—including disease descriptions, symptoms, occurrence conditions, preventive measures, and natural remedies—are converted into dense semantic vector representations using a sentence embedding algorithm. Each textual document s is mapped to a fixed-length embedding vector $\phi(s) \in \mathbb{R}^d$, where semantically similar disease descriptions are positioned closer in the embedding space.

These embeddings are indexed using FAISS (Facebook AI Similarity Search), which enables a fast and scalable similarity search even for large knowledge repositories. During inference, a user query or detected disease name is converted into a query embedding q. FAISS retrieves the top-k most by minimizing the Euclidean distance between the query vector and stored embeddings:

$$i^* = \arg\min_i \|q - v_i\|_2 \quad (12)$$

where v_i represents the stored embedding of the i-th disease document. This approach ensures low-latency retrieval while preserving semantic relevance, making it suitable for real-time advisory systems.

E. Retrieval-Augmented Generation

To ensure that generated advisory responses remain factual, reliable, and explainable, a Retrieval-Augmented Generation (RAG) framework is employed. Instead of generating responses purely from a language model's internal knowledge, the system constrains generation using disease-specific information retrieved through FAISS.

Let x denote the user query or detected disease label and $r \in \mathbf{R}$ represent the retrieved disease documents. The conditional probability of generating a response y is defined as:

$$P(y \mid x) = \sum_{r \in \mathbf{R}} P(y \mid x, r) \, P(r \mid x) \quad (13)$$

This formulation ensures that the generated output is grounded in retrieved factual content, significantly reducing hallucinations and improving interpretability. In the proposed system, the information retrieved from tobacco diseases is passed as a contextual input to the language generation module, which synthesizes coherent and actionable advisory recommendations for farmers.

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

F. Intelligent System Orchestration and Deployment

An intelligent agent is used to coordinate the execution of the proposed system by dynamically selecting processing pathways based on the input modality. Image-based inputs trigger the Vision Transformer–based disease detection module, followed by semantic retrieval and advisory generation, while text-based queries directly invoke the retrieval and response generation components. This adaptive orchestration ensures seamless integration between vision and language modulesss.

The complete framework is deployed as a Flask-based web application, integrating disease detection, semantic embedding, FAISS retrieval, retrieval-augmented generation, and agent logic within a unified backend. Methodologically, the proposed system extends traditional disease classification approaches by providing an end-to-end decision-support framework that combines automated diagnosis with explainable, knowledge-driven recommendations for tobacco disease management.

IV. Experimental Results and Discussion

This section presents the experimental evaluation of the proposed tobacco leaf disease detection system and provides a comparative analysis against baseline deep learning models. The performance of the Vision Transformer (ViT)–based model is evaluated using standard classification metrics and compared with Convolutional Neural Network (CNN), ResNet, and Recurrent Neural Network (RNN) architectures to demonstrate the effectiveness of the proposed approach.

A. Experimental Setup

All models were trained and evaluated on the same tobacco leaf dataset to ensure a fair comparison. The dataset was divided into training, validation, and test sets. Image preprocessing techniques, including resizing to 224 × 224 pixels and normalization, were consistently applied across all models. The following architectures were evaluated:

- Vision Transformer (ViT)
- Convolutional Neural Network (CNN)
- ResNet
- Recurrent Neural Network (RNN)

Model performance was assessed using training accuracy, validation accuracy, test accuracy, loss curves, confusion matrix, and classification metrics

B. Model Comparison Based on Accuracy

Table 1 presents the best validation accuracy and test accuracy achieved by each model.

TABLE I
Performance comparison of different models

Model	Best Validation Accuracy (%)	Test Accuracy (%)
Vision Transformer (ViT)	**99.64**	**88.61**
ResNet	99.60	85.33
CNN	77.73	62.56
RNN	60.40	~60

Discussion: The Vision Transformer achieved the highest validation and test accuracy among all evaluated models. While ResNet demonstrated strong performance, it slightly underperformed compared to ViT in terms of generalization on unseen test data. CNN showed moderate learning capability, whereas RNN performed poorly due to its inability to capture spatial features effectively, confirming that RNNs are unsuitable for image-based disease detection tasks.

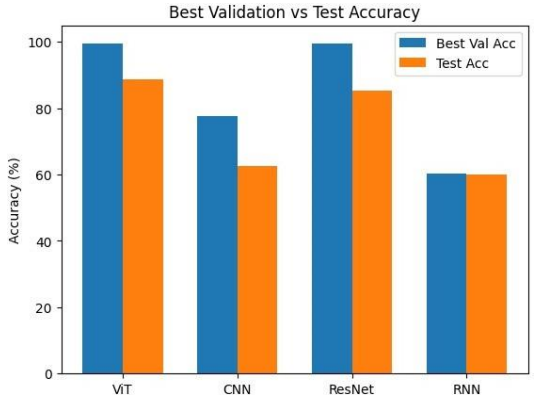

Fig. 3. Comparison of best validation accuracy and test accuracy across different models

C. Training and Validation Accuracy Analysis

Figures illustrating training accuracy vs epoch and validation accuracy vs epoch show the convergence behavior of all models. **Observations**:

- ViT and ResNet converge rapidly within the first few epochs.
- ViT maintains consistently high validation accuracy close to 100%indicating strong generalization.
- CNN exhibits slower convergence and lower accuracy.
- RNN demonstrates unstable and weak learning trends.

Fig. 4. Training accuracy versus epoch for ViT, CNN, ResNet, and RNN models

Discussion: The self-attention mechanism of ViT enables effective modeling of long-range dependencies in leaf images, resulting in faster convergence and superior performance compared to convolution-based and sequential models.

ISSN 2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

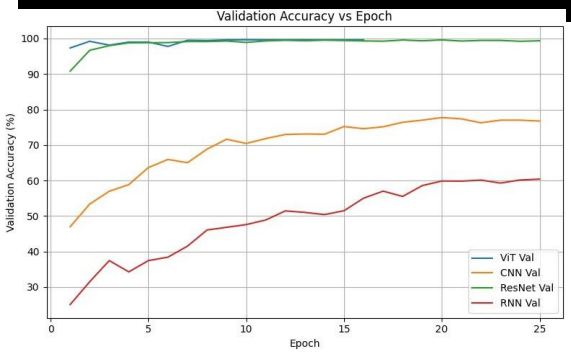

Fig. 5. Validation accuracy versus epoch for ViT, CNN, ResNet, and RNN models

D. Loss Curve Analysis

The training and validation loss curves for the Vision Transformer model are shown in Figure 6. **Discussion:** Both

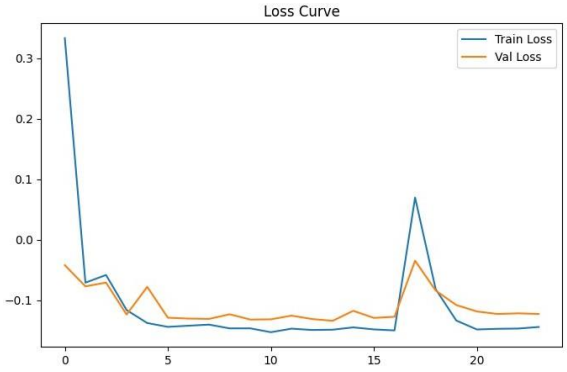

Fig. 6. Training and validation loss curves across epochs

training and validation losses decrease steadily with minimal divergence, indicating stable learning and the absence of overfitting. The slight fluctuations observed during later epochs are attributed to learning rate adjustments and data variability. Overall, the loss behavior confirms the robustness of the proposed model.

E. Confusion Matrix Analysis

The confusion matrix for the Vision Transformer model is presented in Figure 7.

F. Classification Report Analysis

Table 2 summarizes precision, recall, and F1-score for each disease class. **Overall Accuracy:** 88.6%
Macro Average F1-score: 0.87
Weighted Average F1-score: 0.88
Discussion: The classification report confirms balanced performance across disease classes. High F1-scores for most categories indicate effective feature learning and robust classification. The slightly lower precision and recall for a few classes can be attributed to limited inter-class visual variation.

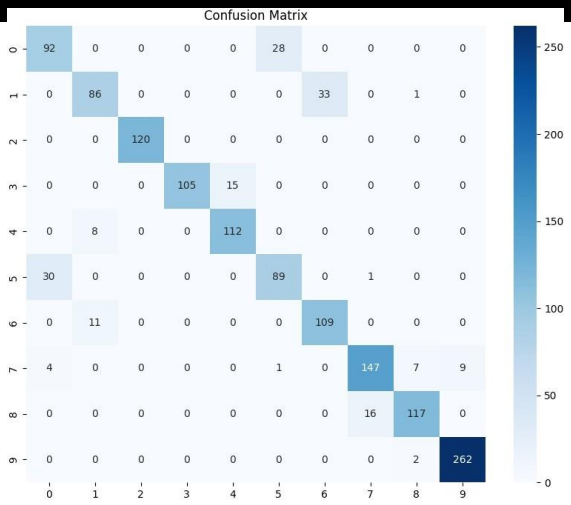

Fig. 7. Confusion matrix illustrating the classification performance of the proposed model

TABLE II
CLASSIFICATION REPORT WITH ACCURACY FOR VISION TRANSFORMER MODEL

Class	Precision	Recall	F1-Score	Accuracy
Class 0	0.73	0.77	0.75	0.77
Class 1	0.82	0.72	0.76	0.72
Class 2	1.00	1.00	1.00	1.00
Class 3	1.00	0.88	0.93	0.88
Class 4	0.88	0.93	0.91	0.93
Class 5	0.75	0.74	0.75	0.74
Class 6	0.77	0.91	0.83	0.91
Class 7	0.90	0.88	0.89	0.88
Class 8	0.92	0.88	0.90	0.88
Class 9	0.97	0.99	0.98	0.99

G. Discussion of Experimental Findings

The experimental evaluation demonstrates that the proposed Vision Transformer consistently outperforms CNN and RNN-based models across all evaluation metrics. The improved performance can be attributed to the model's ability to capture global contextual information through self-attention mechanisms. Unlike convolutional models that focus on localized features, the transformer architecture enables holistic learning of disease patterns distributed across the leaf surface. The relatively lower performance of RNN further confirms that sequential models are not suitable for spatial image analysis tasks. These findings validate the effectiveness of transformer-based architectures for complex agricultural disease detection problems.

V. COMPARATIVE DISCUSSION

The comparative study highlights clear performance differences among the evaluated deep learning architectures. The Vision Transformer demonstrates the strongest generalization capability across validation and test datasets, outperforming CNN, ResNet, and RNN models. This superior performance is attributed to the self-attention mechanism, which enables ef-

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 03(March),2026
Received: 11-02-2026 Accepted: 19-03-2026 Published: 25-03-2026

fective modeling of global spatial dependencies across tobacco leaf images.

ResNet serves as a strong convolutional baseline due to its residual learning framework, which improves gradient flow and stabilizes deep feature learning. Although ResNet achieves high validation accuracy, its reliance on localized convolution operations limits its ability to capture long-range contextual information, resulting in slightly reduced generalization performance when compared to the Vision Transformer.

Conventional CNN models show moderate performance by effectively learning local texture patterns but struggle with complex disease distributions and irregular symptom localization. Recurrent Neural Networks exhibit the weakest performance, confirming their unsuitability for spatial image classification tasks. Overall, these findings validate the adoption of transformer-based architectures for agricultural image analysis and support their integration into precision agriculture decision-support systems.

VI. Key Findings

The experimental evaluation confirms that the Vision Transformer achieved the highest test accuracy of 88.61%, demonstrating superior generalization performance compared to CNN and ResNet architectures. The model exhibited faster convergence and stable loss behavior during training, indicating effective optimization and reduced overfitting. Furthermore, the transformer-based attention mechanism enabled improved discrimination between visually similar disease classes by capturing global spatial dependencies across the leaf surface. These characteristics collectively highlight the effectiveness of attention-driven feature learning and provide strong empirical evidence supporting the adoption of Vision Transformer architectures over conventional CNN-based methods for tobacco leaf disease detection.

VII. Conclusion

This paper proposed an intelligent tobacco leaf disease detection and advisory framework that integrates deep learning–based visual analysis with knowledge-driven information retrieval. A customized Vision Transformer (DIEC-ViT) was employed for disease classification using a tobacco-specific image dataset and demonstrated superior performance over CNN, ResNet, and RNN models in terms of accuracy and generalization.

To enhance practical usability, a structured tobacco disease knowledge base was developed and semantically indexed using sentence embeddings and FAISS for efficient retrieval. The incorporation of a retrieval-augmented generation mechanism ensured that advisory responses were factual, reliable, and explainable. The system was deployed as a web-based application, validating its feasibility for real-world agricultural use.

Overall, the proposed framework effectively combines disease classification, semantic retrieval, and advisory intelligence within a unified architecture, highlighting the potential of transformer-based models and semantic search techniques for precision agriculture applications.

VIII. Limitations and Future Work

Although the proposed tobacco leaf disease detection and advisory system demonstrates strong performance and practical applicability, certain limitations remain. The tobacco leaf image dataset used in this study is limited in size and primarily consists of images captured under controlled environmental conditions. As a result, the model's performance under real-field conditions with varying illumination, background noise, and occlusions has not been fully explored.

Additionally, the current framework is designed to predict a single disease class per input image. In real agricultural scenarios, multiple diseases may coexist on the same leaf, which is not addressed in the present implementation. Furthermore, disease severity estimation, which could provide more precise treatment recommendations, is beyond the scope of this study.

Future work will focus on expanding the dataset with large-scale, real-field images collected across different growth stages and environmental conditions to improve model robustness and generalization. The framework can also be extended to support multi-disease detection and severity grading. Optimization for mobile and edge-device deployment will be explored to enable real-time, on-field diagnosis. Moreover, integrating multilingual advisory support and explainable visualization techniques, such as attention heatmaps, can further enhance system transparency, accessibility, and user trust.

References

[1] R. Pydipati, T. Burks, and W. Lee, "Identification of citrus disease using color texture features and discriminant analysis," *Computers and Electronics in Agriculture*, 2006.
[2] S. P. Mohanty, D. P. Hughes, and M. Salathe´, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, 2016.
[3] Y. Zhang, S. Wang, and G. Ji, "Tobacco disease recognition based on convolutional neural networks," *IEEE Access*, 2019.
[4] K. Lin, J. Gong, and H. Li, "Deep learning-based classification of tobacco leaf diseases," *Computers and Electronics in Agriculture*, 2020.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[6] J. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, 2019.
[7] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021.
[8] M. Yuan, Y. Chen, and Z. Wang, "Vision transformer-based plant disease recognition," *IEEE Access*, 2022.
[9] S. Li, C. Zhang, and J. Wang, "Attention-based transformer networks for plant leaf disease detection," *Computers and Electronics in Agriculture*, 2022.
[10] T. Mikolov *et al.*, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.
[11] J. Johnson, M. Douze, and J. Je´gou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, 2019.
[12] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[13] G. Keerthi, G. Ramachandran, and G. V. S. N. R. V. Prasad, "Cuckoo search optimization-based feature selection for predicting autism spectrum disorder using artificial immune algorithms," *Journal of Theoretical and Applied Information Technology*, Jan. 2025.

[14] G. Keerthi, G. Ramachandran, and G. V. S. N. R. V. Prasad, "Autism spectrum disorder prediction using LASSO-regularised bat search optimisation," *International Journal of Services Operations and Informatics*, 2024.
[15] G. Keerthi, Y. Ashvitha, V. S. R. Reddy, R. M. Krishna, and P. Sandeep, "Integrating convolutional neural networks and machine learning for accurate identification of autism spectrum disorder using facial biomarkers," in *Proc. IEEE Int. Conf. on Emerging Systems and Intelligent Computing (ESIC)*, Feb. 2024.
[16] M. Babu Rao, Ch. Kavitha, B. Prabhakara Rao, and A. Govardhan, "A new feature set for content based image retrieval," in *Proc. IEEE Sponsored Int. Conf. on Information, Communication and Embedded Systems (ICICES)*, Feb. 2013.
[17] P. Ramya, V. Ramya, and M. Babu Rao, "E-waste management using hybrid optimization-enabled deep learning in IoT-cloud platform," *Advances in Engineering Software*, vol. 176, Art. no. 103353, 2023.