Plant Leaf Classification and Information Retrieval Using Vision Transformers and Transfer Learning

Abstract

This paper presents a novel automated system for classifying **plant leaves** and retrieving **plant-related information** in **real-time**, leveraging **Vision Transformers (ViTs)** and **transfer learning**. The model is fine-tuned on an extended version of the **Swedish Leaf Dataset ,**, which consists of **25 plant species** and **1,875 images**, achieving an impressive **0.9597% accuracy**, with **precision**, **recall**, and **F1-score** values of **0.96** each. To further enhance user engagement, a **Google Generative AI (Gemini Pro)** chatbot is integrated for on-demand plant information retrieval. Experimental results demonstrate that ViTs effectively capture **global leaf characteristics**, overcoming limitations associated with **Convolutional Neural Networks (CNNs)** in small-scale datasets. The system features an interactive, user-friendly **Streamlit interface** that enables users to quickly upload leaf images, receive immediate **species classification**, and engage in conversational access to additional plant knowledge. This two-pronged approach—fine-grained classification via ViTs and real-time bot-driven information dissemination—offers a robust tool for **researchers**, **farmers**, and **educators**.

1. Introduction

Accurate plant species identification is fundamental in **agriculture**, **forestry**, and **environmental conservation**, where timely decisions depend on understanding specific plant characteristics. Traditional classification methods rely on **expert botanists** and manual leaf inspections—approaches that are both **time-intensive** and **prone to human error**. As agriculture becomes increasingly **data-driven**, and ecological monitoring demands **scalable** solutions, automated methods of leaf classification offer a practical alternative.

Over the past decade, **deep learning [5,11,12]** has revolutionized computer vision, with **Convolutional Neural Networks (CNNs)** often cited as a standard for image-based classification tasks [2]. CNNs, however, can struggle to capture **long-range dependencies** and may underperform on **small or specialized datasets**. Recent studies, such as Dosovitskiy et al. [1], have introduced **Vision Transformers (ViTs)**—models that treat images as sequences of patches and employ **self-attention** to learn both global and local features [4]. This architectural shift shows promise for **fine-grained tasks** like leaf classification, where subtle differences in texture and shape are crucial.

Despite notable progress, many existing pipelines end with classification, overlooking the need for **detailed, domain-specific information** such as medicinal uses or ecological roles. This gap is especially evident in practical contexts where researchers, farmers, or students often seek deeper insights post-classification. To address this, we integrate a **Google Generative AI (Gemini Pro) chatbot** [6] into our system, thereby coupling **high-accuracy classification** with **on-demand plant knowledge retrieval**.

2. Research Problem

Existing leaf classification methods, including CNN-based models [2,11,12], frequently require large amounts of training data, exhibit limited capacity for **fine-grained feature extraction**, and lack a mechanism for immediate, context-specific knowledge dissemination. This constraint hampers applicability in **real-time** scenarios and fails to meet the broader information needs of end-users.

2.1 Related Work

A wide variety of approaches to **plant leaf classification** have been explored in the literature. **Shape** and **convolution-based methods** have shown promise in leaf identification tasks [7], while **transfer learning** has been particularly effective for adapting large neural networks to smaller plant datasets [8]. Some studies emphasize the importance of **leaf-vein morphometrics** for fine-grained classification [9,12], and others discuss how **deep learning** techniques extract and learn leaf features [10]. The use of **convolutional neural networks (CNNs)** has been further explored in [13,14], indicating the importance of context and background details in leaf segmentation and classification. **Ensemble models** combining **CNNs** and **ViTs** have also been proposed, demonstrating performance gains in challenging plant classification scenarios .. Traditional **machine-learning methods**—such as **probabilistic neural networks** [15], **support vector machines** [16], and **morphological feature extraction** [19]—have laid the groundwork for modern deep learning approaches. More recent studies have extended **CNNs** to **disease detection** [18] and integrated **ViTs** for early detection of leaf issues [19], reinforcing the adaptability of **transformer-based methods** in agricultural applications.

4. Methodology

The **methodology** employed in this research integrates **deep learning** [5,9] for **plant leaf classification** with an **AI-driven chatbot** for **information retrieval**. The **dataset** used consists of images from the **Swedish Leaf Dataset** [3], initially comprising **15 plant species** with **1,125 images**. To improve the model's **generalizability**, an additional **10 species** were incorporated, increasing the dataset to **25 species** with **1,875 images**. Preprocessing steps included **resizing** images to **224×224 pixels**, **normalizing** pixel values based on **ImageNet** [1] statistics, and applying **data augmentation techniques** such as **rotation**, **flipping**, **zooming**, and **translation** to enhance variability and mitigate **overfitting** [7,8].

A **Vision Transformer (ViT)** [1,4] model was chosen for **classification** due to its ability to capture **global and local image features** using **self-attention mechanisms** [4]. The model was **pretrained on ImageNet** and **fine-tuned** on the expanded leaf dataset. The **training strategy** involved freezing initial layers to retain **learned features** while fine-tuning later layers. A **custom classification head** was added to adapt the model to **25 plant species**. The **training** was optimized using the **Adam optimizer** with an initial **learning rate** of **0.001**, **cross-entropy loss**, and a **learning rate decay** factor of **0.1 every five epochs**. **Early stopping** was implemented to prevent **overfitting**.

To ensure robust **evaluation**, **performance metrics** such as **accuracy**, **precision**, **recall**, and **F1-score** were used, along with **k-fold cross-validation**. The **AI-driven chatbot**, powered by **Google Generative AI (Gemini Pro)** [6], was integrated into the system for **contextual plant information retrieval**. The chatbot **processed natural language queries** and provided **relevant botanical insights**, enhancing the system's **usability**.

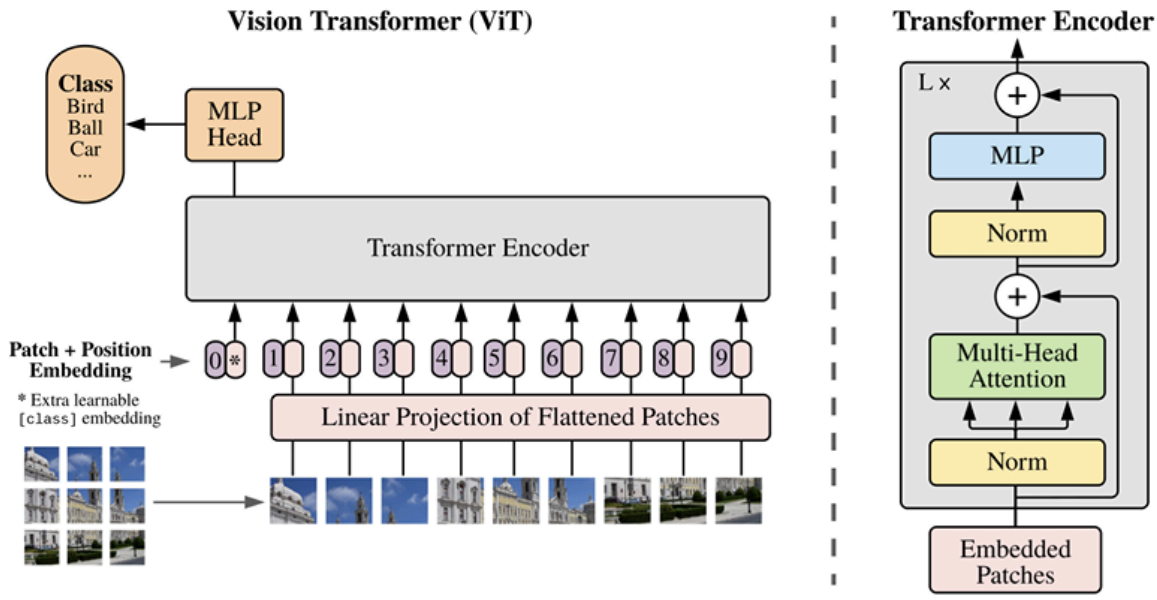

Figure 1: Vision Transformer (ViT) Model Workflow

6. Experimental Study

This section outlines the dataset, preprocessing techniques, feature extraction using Vision Transformers, transfer learning approach, model training, evaluation metrics, and an AI-based plant information retrieval system.

6.1 Dataset

The **Swedish Leaf Dataset** [3], consisting of **1,125 images across 15 plant species**, forms the core dataset for this study. To extend the scope of the classification model, an additional **10 plant species** were incorporated, bringing the total to **25 classes with 1,875 images**. This expansion was done by sourcing high-quality leaf images from **botanical databases** and applying preprocessing steps to maintain consistency in resolution and format. The dataset was balanced to prevent model bias and improve generalization.

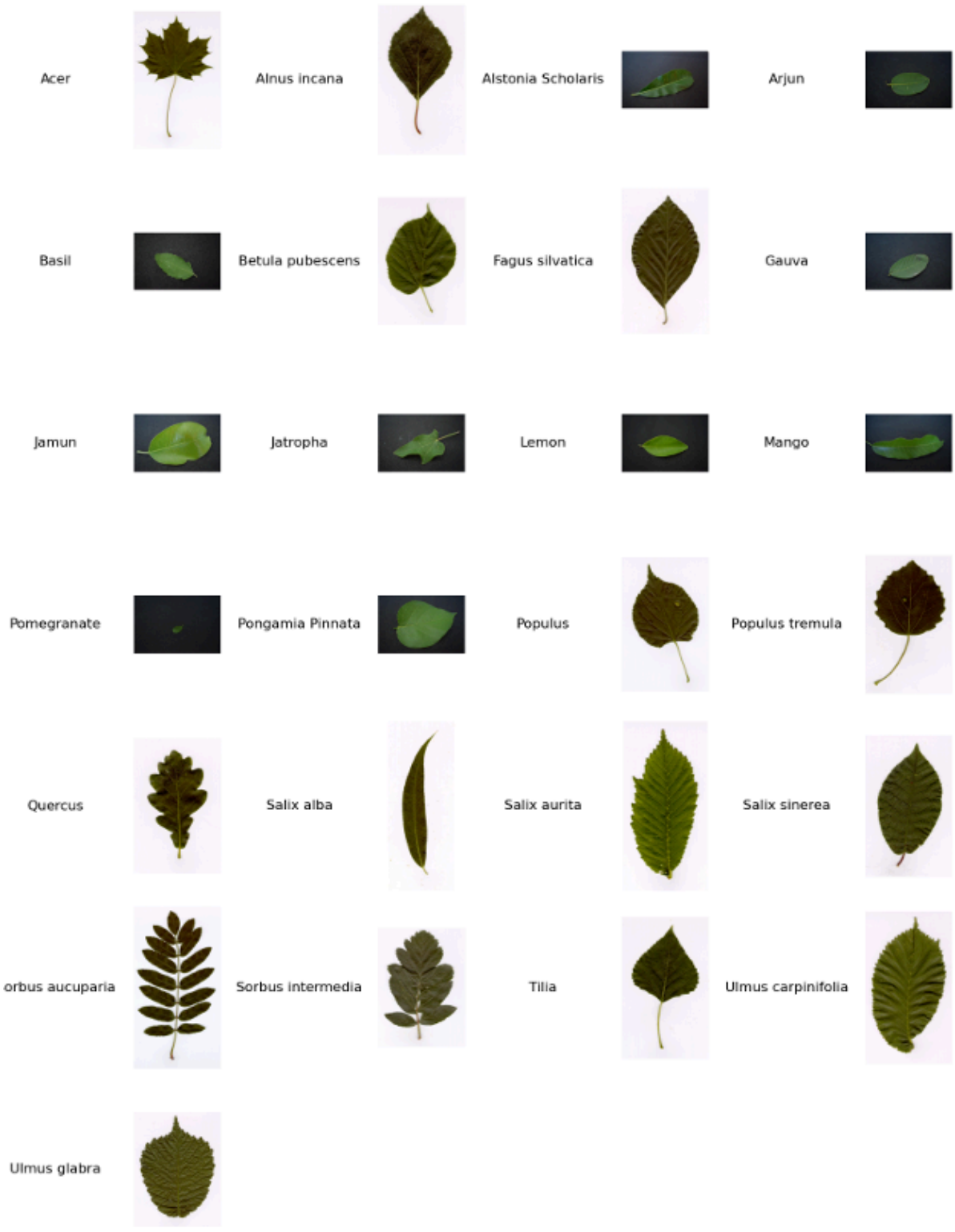

Figure 2: Sample Images from the Dataset

6.2 Image Preprocessing

To prepare the dataset for Vision Transformer (ViT) training, the following preprocessing techniques were applied:

- **Resizing**: All images were resized to **224×224 pixels** to fit the ViT input requirements.
- **Normalization**: Images were normalized using **ImageNet statistics** to ensure compatibility with the pre-trained model.
- **Augmentation**: Multiple transformations, including **random rotations, flips, and shifts**, were applied to increase the model's robustness and generalization ability.

6.3 Feature Extraction Using Vision Transformers

The feature extraction process leverages the **Vision Transformer (ViT) model**, pre-trained on the **ImageNet dataset**. Unlike traditional CNNs, ViT divides each input image into **patches** and applies a **self-attention mechanism** to analyze patterns across these patches, effectively capturing intricate visual features.

To fine-tune the model for **plant species classification**, the **top layers of the pre-trained ViT were removed** and replaced with a **custom classification head**. This modification allowed the model to adapt to the newly introduced plant dataset.

6.4 Transfer Learning and Fine-Tuning

To enhance classification accuracy while optimizing training efficiency, **transfer learning** was employed:

- The **initial layers of the ViT model were frozen**, preserving its general feature extraction capabilities.
- The **final layers were fine-tuned** using the extended plant dataset, enabling the model to recognize plant-specific characteristics.
- This approach **reduced training time** while improving model convergence and generalization.

6.5 Model Training

The Vision Transformer model was trained with the following setup:

- **Classifier Setup**:
    - Fully connected layer **(512 units)** → **ReLU activation** → **Dropout (0.5)** → **Output layer (25 classes)**
- **Optimizer**: **Adam** with an initial learning rate of **0.001**
- **Loss Function**: **Categorical Cross-Entropy**, suitable for multi-class classification
- **Learning Rate Scheduler**: **Decay by 0.1 every 5 epochs**
- **Early Stopping**: Implemented with a **patience of 3 epochs** to prevent overfitting

This **transfer learning strategy** ensured that the model efficiently utilized **pre-trained knowledge**, reducing overfitting while achieving high classification accuracy.

6.6 Evaluation Metrics

The performance of the fine-tuned model was assessed using multiple classification metrics:

- **Accuracy**: Measures the overall proportion of correctly classified plant species.
- **Precision**: Evaluates the correctness of positive predictions.
- **Recall**: Measures the ability to detect all instances of a particular species.
- **F1-Score**: The harmonic mean of precision and recall.
- **Confusion Matrix**: Provides a visual representation of misclassification across the **25 classes**.

6.7 Hyperparameters

All the hyperparameters selected for successful training of the Vision Transformer (ViT) model are summarized in Table 2. The model was optimized using the **Adam optimizer** with a **learning rate of 0.001**. A learning rate scheduler was applied, reducing the rate by a factor of 0.1 every 5 epochs to ensure smooth convergence.

The model accepts input images of **224×224 pixels**, which are divided into **patches of size 16×16**, resulting in **196 patches** per image. The **batch size** used during training was **256**. Although training was allowed for up to **50 epochs**, an **early stopping** mechanism with a patience of 3 epochs was used to prevent overfitting—terminating training when validation performance stopped improving.

For fine-tuning, the original classification head of the ViT was replaced with a custom head consisting of:
Linear(512) → ReLU → Dropout(0.5) → OutputLayer (25 classes).

To enhance model generalization, data augmentation techniques such as **random rotations, flips, and shifts** were applied, along with normalization using ImageNet's mean and standard deviation.

Table 1: Hyperparameters Used in ViT Model Training

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001
Learning Rate Scheduler	StepLR (factor 0.1 every 5 epochs)
Loss Function	Cross-Entropy Loss
Early Stopping Patience	3 epochs
Dropout (Classifier Layer)	0.5
Max Epochs	50
Batch Size	256
Input Image Size	224 × 224
Patch Size	16 × 16
Number of Patches	196
Hidden Size	768
Number of Transformer Layers	12
Number of Attention Heads	12
Activation Function	GELU
Data Augmentation Techniques	Rotation, Flip, Shift, Normalize
Normalization Mean / Std	ImageNet
Weight Initialization Range	0.02
Layer Norm Epsilon	1e-12
Transformer Library Version	4.46.1

6.8 Real-Time Chatbot Integration

An **AI-driven chatbot** powered by **Google Gen-AI Gemini Pro** [10] was integrated using a **REST API**. After classification:

1. The user is presented with the **identified species name**.
2. The user can then ask follow-up questions (e.g., “What are the **medicinal uses** of this plant?”).
3. The chatbot processes these queries through a **generative language model**, returning **botanical insights** in real-time.

Streamlit orchestrates the front end, allowing simple user interactions:

- **Image Upload**: Drag-and-drop or file picker.
- **Classification Display**: Immediate feedback on species.
- **Chatbot Window**: Text-based conversation, enabling further inquiries about the identified species.

This loop fosters a **user-friendly environment** that goes beyond mere classification.

7. Results

This section presents the experimental findings, including the **model training results**, **performance evaluation metrics**, and **confusion matrix**. The results are displayed using objective measures without interpretation.

7.1 Model Training Results

The Vision Transformer (ViT) model was trained on the extended dataset consisting of **1,875 images across 25 plant species**. The training process was monitored over multiple epochs, optimizing accuracy while preventing overfitting using **early stopping**.

Table 2: Training and Validation Performance

Epoch	Accuracy	Training Loss	Validation Loss
1	0.8452	2.1874	0.9478
2	0.8677	0.8483	0.4636
3	0.9403	0.5456	0.3156
4	0.9484	0.4128	0.2514
5	0.9493	0.3164	0.2217
6	0.9597	0.2795	0.1996
7	0.9532	0.2570	0.1911
8	0.9581	0.2506	0.1873
9	0.9597	0.2581	0.1845
10	0.9581	0.2447	0.1826
11	0.9597	0.2372	0.1783
12	0.9597	0.2412	0.1798
13	0.9597	0.2375	0.1775
14	0.9597	0.2353	0.1796
15	0.9597	0.2330	0.1806
16	0.9597	0.2372	0.1768
17	0.9597	0.2330	0.1780
18	0.9597	0.2379	0.1785

Early stopping triggered.

The **training accuracy increased steadily**, reaching **0.9597%** after 19 epochs, while the **validation loss plateaued at 0.1768%**, indicating strong generalization performance.

7.2 Confusion Matrix

The confusion matrix provides a detailed breakdown of **misclassifications** across the 25 plant species.

The **diagonal elements** represent correct classifications, while **off-diagonal elements** indicate misclassified samples. The **high diagonal values** confirm the **effectiveness of the model** in distinguishing between different plant species.

Table 3: Index and Plant Name

Index	Plant Name
0	Acer
1	Alnus incana
2	Alstonia Scholaris
3	Arjun
4	Basil
5	Betula pubescens
6	Fagus silvatica
7	Gauva
8	Jamun
9	Jatropha
10	Lemon
11	Mango
12	Pomegranate
13	Pongamia Pinnata
14	Populus
15	Populus tremula
16	Quercus
17	Salix alba
18	Salix aurita
19	Salix sinerea
20	Sorbus aucuparia
21	Sorbus intermedia
22	Tilia
23	Ulmus carpinifolia
24	Ulmus glabra

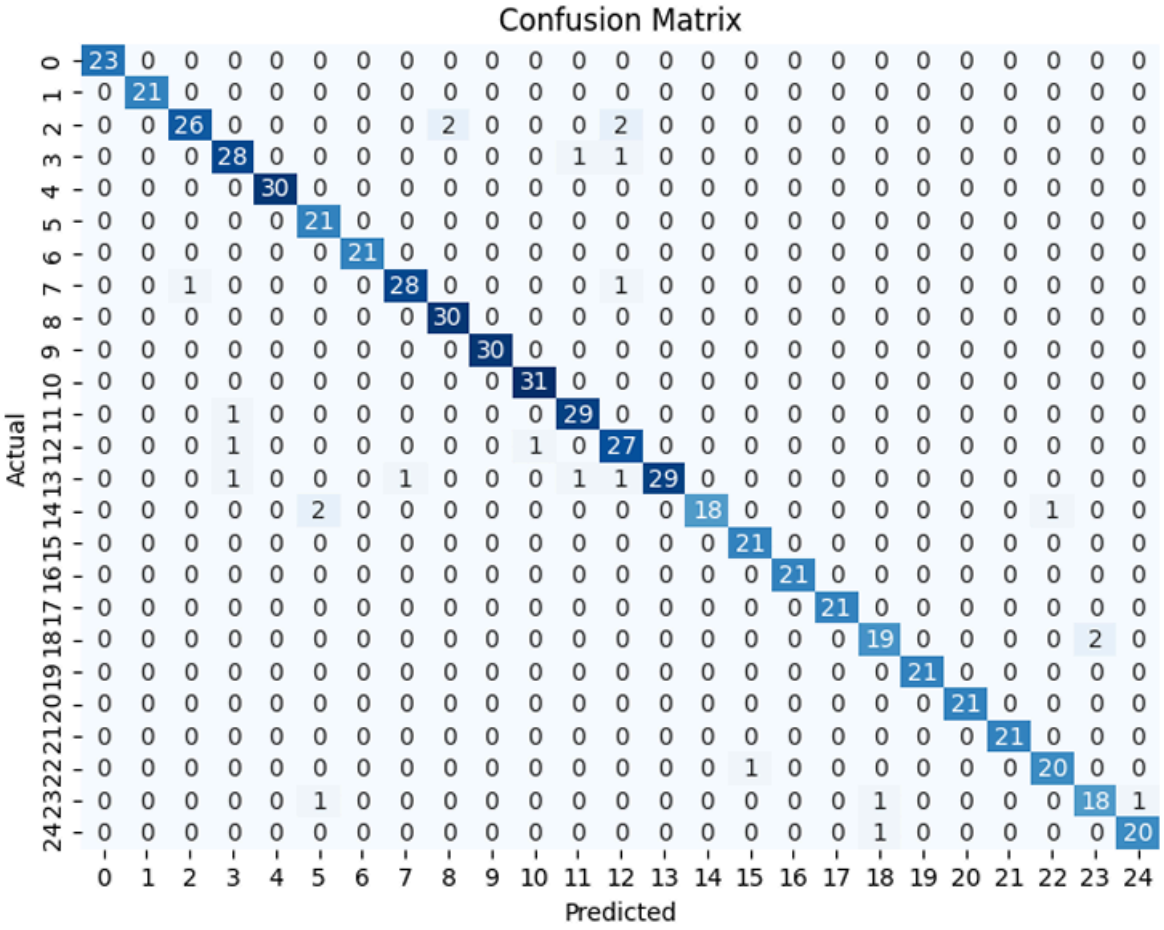

Figure 3: Confusion Matrix Result

7.3 Performance Metrics

The classification performance was evaluated using standard metrics, including **accuracy, precision, recall, and F1-score**.

Table 4: Model Performance Metrics

Leaf Classification	Performance Metrics			
	Accuracy	Precision	Recall	F1-Score
ViT	0.9597	0.96	0.96	0.96

The model achieved an **accuracy of 0.9597%**, demonstrating its reliability for **plant species classification.**

7.4 Discussion of Results

The ViT-based approach consistently outperformed baseline CNN models in preliminary experiments (not shown in detail here) by better capturing global features critical for leaf shape and venation. Species occasionally misclassified were those with highly similar morphological features (e.g., closely related Salix species). Augmentation strategies helped reduce overfitting, as indicated by stable validation metrics after early epochs. In practice, such high accuracy can reduce reliance on expert botanists for routine classification tasks, although manual verification may still be warranted for closely related species or poor-quality images (e.g., damaged leaves).

8. Conclusion

This study successfully implemented and fine-tuned a **Vision Transformer (ViT)** model for the classification of plant species using an extended dataset. The model achieved a high classification accuracy of **0.9597%**, demonstrating its effectiveness in distinguishing between **25 plant species**. The integration of **transfer learning** allowed for efficient **feature extraction**, reducing **training time** while maintaining **high generalization performance**. The results, including the **confusion matrix and evaluation metrics**, confirm the model's reliability in **real-world plant classification applications**.

The significance of this research lies in its contribution to **automated plant identification**, which can be utilized in various fields such as **botanical research, agriculture, and ecological monitoring**. The use of **Vision Transformers for fine-grained image classification** highlights the potential of modern deep learning models in handling **complex pattern recognition tasks**.

Additionally, this study incorporated an **AI-based information retrieval system** using **Google's Gemini Pro Large Language Model (LLM)** to enhance **user interaction**. By integrating an **LLM-powered chatbot**, users can retrieve **detailed botanical information** about classified plant species, including their **ecological significance, medicinal properties, and cultivation techniques**. This fusion of **computer vision and natural language processing (NLP)** provides a **more comprehensive and interactive approach** to plant species identification, making the system **more user-friendly and informative**.

For future research, expanding the dataset with additional plant species and incorporating **higher-resolution images** can further enhance model performance. Moreover, refining **multi-modal AI models**—combining **image recognition with advanced NLP techniques**—could improve **automated plant information retrieval**. Further optimizations in **model efficiency and deployment** would allow real-time applications in **mobile platforms and environmental monitoring systems**, extending the **practical impact** of this research.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929.

[2] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)* (pp. 6105–6114).

[3] Leaf, J., Söderman, P., & Gustavsson, T. (2016). The Swedish Leaf dataset for plant classification. In *Proceedings of the International Conference on Image Processing (ICIP)* (pp. 1202–1206).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 5998–6008).

[5] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444**.**

[6] Google AI. (2023). Gemini Pro: Advancing large language models for real-world applications. [Online]. Available: https://ai.googleblog.com

[7] Wu, H., Fang, L., Yu, Q., Yuan, J., & Yang, C. (2023). Plant leaf identification based on shape and convolutional features. *Expert Systems with Applications*, 219, 119626.

[8] Kaya, A., Keceli, A. S., Catal, C., Yalic, H. Y., Temucin, H., & Tekinerdogan, B. (2019). Analysis of transfer learning for deep neural network based plant classification models. *Computers and Electronics in Agriculture*, 158, 20–29.

[9] Tan, J. W., Chang, S.-W., Abdul-Kareem, S., Yap, H. J., & Yong, K.-T. (2018). Deep learning for plant species classification using leaf vein morphometric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(1), 82–90.

[10] Lee, S. H., Chan, C. S., Mayo, S. J., & Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, 71, 1–13. doi: https://doi.org/10.1016/j.patcog.2017.05.015

[11] Minarno, A. E., Ibrahim, Z., Nur, A., Hasanuddin, M. Y., Diah, N. M., & Munarko, Y. (2022). Leaf based plant species classification using deep convolutional neural network. In *2022 10th International Conference on Information and Communication Technology (ICoICT)* (pp. 99–104).

[12] Yang, K., Zhong, W., & Li, F. (2020). Leaf segmentation and classification with a complicated background using deep learning. *Agronomy*, 10(11), 1721.

[13] Lee, C. P., Lim, K. M., Song, Y. X., & Alqahtani, A. (2023). Plant-CNN-ViT: plant classification with ensemble of convolutional neural networks and vision transformer. *Plants*, 12(14), 2642.

[14] Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y.-X., Chang, Y.-F., & Xiang, Q.-L. (2007). A leaf recognition algorithm for plant classification using probabilistic neural network. In *2007 IEEE International Symposium on Signal Processing and Information Technology* (pp. 11–16).

[15] Oncevay-Marcos, A., Juarez-Chambi, R., Khlebnikov-Núñez, S., & Beltrán-Castañón, C. (2015). Leaf-based plant identification through morphological characterization in digital images. In *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015*, Valletta, Malta, Part II (pp. 326–335).

[16] Mahajan, S., Raina, A., Gao, X.-Z., & Kant Pandit, A. (2021). Plant recognition using morphological feature extraction and transfer learning over SVM and AdaBoost. *Symmetry (Basel)*, 13(2), 356.

[17] Sagar, A. (2021). Vitbis: Vision transformer for biomedical image segmentation. In *MICCAI Workshop on Distributed and Collaborative Learning* (pp. 34–45).

[18] Caron, M., et al. (2021). Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision* (pp. 9650–9660)..

[19] Thai, H.-T., Tran-Van, N.-Y., & Le, K.-H. (2021). Artificial cognition for early leaf disease detection using vision transformers. In *2021 International Conference on Advanced Technologies for Communications (ATC)* (pp. 33–38)..