

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



**Manish Kumar**

Advisor: **Dr. Subhasish Dhal**

Department of Computer Science and Engineering  
Indian Institute of Information Technology Guwahati

This dissertation is submitted for the degree of  
*Bachelors of Technology*

*to my loving parents...*

## **Declaration**

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Manish Kumar  
November 2024

## **Acknowledgements**

I acknowledge Dr. Subhasish Dhal for the invaluable guidance and support provided throughout this project. I am also grateful to the faculty members of the CSE Department at IIITG for providing me with a stimulating academic environment and the opportunity to pursue this research. I extend my thanks to facultites, whose courses and discussions have broadened my understanding of the subject matter.

## **Abstract**

This thesis presents an automated plant species classification system using Vision Transformers (ViTs) combined with transfer learning and AI, aiming to replace manual identification with accurate, automated leaf image analysis. The report discusses background, motivation, and research objectives, focusing on fine-tuning ViTs for classifying plant species on an extended custom dataset and integrating a Google Gen-AI chatbot (Gemini Pro) for real-time plant information. A literature review compares existing malware classification techniques, noting limitations of traditional methods. The proposed model applies data preprocessing, feature extraction, and transfer learning on the Swedish Leaf Dataset, enhanced with additional species. Model training utilized early stopping, learning rate scheduling, and cross-entropy loss, with evaluation metrics such as precision, recall, F1-score, and accuracy. Results show ViTs excel at capturing leaf details, outperforming CNNs, with high classification accuracy. Future work includes adding multilingual support, disease detection, and scalability to other datasets for enhanced versatility.

# Table of Contents

<b>List of Figures</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Motivation . . . . .	1
1.3 Research Objectives . . . . .	2
1.4 Innovation through Transfer Learning . . . . .	2
1.5 Contributions . . . . .	2
<b>2 Related Work in Plant Leaf Classification</b>	<b>4</b>
2.1 Literature Survey . . . . .	4
2.2 Limitations and Challenges in Existing Approaches . . . . .	5
2.3 Summary . . . . .	6
<b>3 Proposed Model</b>	<b>7</b>
3.1 Overview . . . . .	7
3.2 Dataset . . . . .	7
3.3 Image Preprocessing . . . . .	8
3.4 Feature Extraction Using Vision Transformers . . . . .	8
3.5 Transfer Learning and Fine-Tuning . . . . .	8
3.6 Model Training . . . . .	9
3.7 Evaluation Metrics . . . . .	9
3.8 Plant Information Retrieval using Google Gen-AI . . . . .	9
3.9 Computational Environment . . . . .	10
3.10 Vision Transformer (ViT) Architecture . . . . .	10
<b>4 Results</b>	<b>13</b>
4.1 Confusion Matrix Result . . . . .	13
4.2 Model Training Result . . . . .	14

**Table of Contents**

---

4.3	Other Results . . . . .	15
4.4	Conclusion . . . . .	15
4.5	Future Work . . . . .	15
	<b>References</b>	<b>16</b>

# List of Figures

3.1	Sample images from the expanded Swedish Leaf Dataset.[3]	11
3.2	Vision Transformer architecture showcasing patch embeddings and transformer layers. [1]	12
4.1	Confusion Matrix	13
4.2	Transformer Model Training	14
4.3	Other Result	15



# Chapter 1

## Introduction

### 1.1 Background

Plants are essential to ecosystems, agriculture, forestry, and human well-being, yet identifying species accurately remains challenging. Traditional methods rely on manual expert analysis, which is time-consuming, error-prone, and labor-intensive. Existing tools often lack specificity, leading to broad and sometimes inaccurate results. Recent advances in deep learning and computer vision offer automated solutions that are faster, more reliable, and scalable. Automating plant identification not only improves productivity in agriculture and forestry but also enables rapid decision-making in critical situations, such as managing crop disease outbreaks, while contributing to environmental conservation and biodiversity research.

### 1.2 Motivation

The success of Vision Transformers (ViTs) in image classification tasks has been groundbreaking, achieving state-of-the-art results in domains traditionally dominated by Convolutional Neural Networks (CNNs). Vision Transformers have shown exceptional performance in capturing intricate details from images, making them well-suited for tasks that require fine-grained classification, such as distinguishing between similar plant species based on leaf patterns. Additionally, by leveraging transfer learning from models pre-trained on large datasets like ImageNet, we can significantly reduce the data requirements for training plant-specific classifiers. The integration of an AI-driven chatbot powered by the Google Generative AI Gemini

Pro LLM model further enhances the project, allowing users to interactively query information about identified plants.

### 1.3 Research Objectives

This project aims to:

- Utilize Vision Transformers for classifying plant species from leaf images.
- Apply transfer learning to fine-tune pre-trained models, reducing training time and data needs.
- Develop a Streamlit application for users to upload images, receive predictions, and interact with a Gemini Pro-powered chatbot.
- Assess model performance using metrics like accuracy, precision, recall, and F1-score.

### 1.4 Innovation through Transfer Learning

This project innovates by applying transfer learning techniques to fine-tune pre-trained Vision Transformer models (ViT) specifically for the domain of plant leaf classification. By freezing initial layers that capture general image features and fine-tuning the later layers, the model can adapt to the unique characteristics of plant leaf images. This fine-tuning process not only enhances classification accuracy but also improves robustness in distinguishing between visually similar species.

### 1.5 Contributions

The contributions of this thesis are twofold:

- **Methodological Innovation:** This project leverages Vision Transformer (ViT) models, applying transfer learning techniques to fine-tune these pre-trained models for plant leaf classification. It introduces a novel approach by integrating an AI-powered chatbot using the Google Generative AI Gemini Pro LLM model for information retrieval, enhancing user interaction and engagement.

- **Enhanced Feature Extraction:** The fine-tuning of Vision Transformers using transfer learning significantly improves their discriminative capabilities, specifically tailored to capture the subtle visual nuances in plant leaf images. This allows the model to achieve high classification accuracy even with limited domain-specific data.

## Chapter 2

# Related Work in Plant Leaf Classification

### 2.1 Literature Survey

#### Alexey Dosovitskiy et al. [1], 2021

- This study explores the application of Vision Transformers (ViTs) for image classification tasks, traditionally dominated by convolutional neural networks (CNNs).
- ViTs process input images as sequences of patches, similar to tokens in natural language processing, allowing them to achieve competitive performance without the inductive biases present in CNNs.
- The research demonstrates that ViTs can outperform state-of-the-art CNNs like ResNets when pre-trained on large datasets (e.g., ImageNet-21k and JFT-300M).
- The model achieves high accuracy on various benchmarks, including 88.55% on ImageNet and 94.55% on CIFAR-100, with significantly reduced computational resources.

#### Surleen Kaur<sup>1</sup>, Prabhpreet Kaur<sup>2</sup> [3], 2019

- Their study focuses on automated plant species identification using plant leaf images, employing Computer Vision and machine learning techniques.

---

## 2.2 Limitations and Challenges in Existing Approaches

- The system involves four main steps: image acquisition, pre-processing, feature extraction, and classification using Multiclass-Support Vector Machine (MSVM).
- The Swedish Leaf Dataset, containing 1,125 images of 15 species, was used, achieving a classification accuracy of 93.26%.

### **Kiran S.Gawli, Ashwini S. Gaikwad [2], 2020**

- The paper presents an automated plant species classification system using a Convolutional Neural Network (CNN).
- The model includes multiple layers, such as convolutional layers for extracting detailed features, pooling layers to reduce dimensionality, and fully connected layers for the final classification.
- The process involves image pre-processing , feature extraction (texture and color), and classification with a CNN implemented using TensorFlow.
- The model was trained on a dataset of 17 plant species, achieving an accuracy of 94.26%.

## 2.2 Limitations and Challenges in Existing Approaches

- It uses a dataset of limited number of plant species, which restricts the model's ability to generalize effectively across a larger, more diverse range of plants.
- The dataset appears to be captured under controlled conditions, which might not accurately reflect the challenges of real-world environments, like different lighting and backgrounds.
- The research does not focus on real-time classification, which could limit the usability of the system in practical scenarios where immediate results are required.
- It lacks comparison to newer deep learning models, which could provide better performance and robustness.

## 2.3 Summary

The literature reviewed in this chapter underscores the advancements in applying computer vision and machine learning techniques for automated plant species identification. The studies demonstrate the effectiveness of feature extraction methods and classifiers like Multiclass Support Vector Machines (MSVMs) and Convolutional Neural Networks (CNNs) in achieving high classification accuracy. However, the limitations related to small, controlled datasets, absence of real-world variability, and lack of real-time processing capabilities highlight the need for further research. These findings provide a solid basis for exploring more advanced models and methodologies to overcome the current challenges in automated plant classification.

# Chapter 3

## Proposed Model

### 3.1 Overview

This chapter outlines the methodology adopted for the classification of plant species using a deep learning-based approach. The process integrates the classification of plant leaf images using a pre-trained Vision Transformer (ViT) model and fine-tuning through transfer learning to enhance classification accuracy. The system also includes a chatbot component powered by Google Generative AI Gemini Pro LLM for information retrieval, providing additional insights about the identified plant species.

### 3.2 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 set of 10 custom plant species was incorporated, bringing the total to 25 classes with a total of 1,875 images. This dataset was expanded with images sourced from botanical databases and processed to ensure consistency in image quality and resolution. The combined dataset is relatively balanced, reducing the risk of model bias.

## 3.3 Image Preprocessing

The images were preprocessed using standard transformations:

- **Resizing:** All images were resized to 224×224 pixels to fit the input requirements of the Vision Transformer.
- **Normalization:** Images were normalized using ImageNet statistics to ensure compatibility with the pre-trained model.
- **Augmentation:** Data augmentation techniques such as random rotations, flips, and shifts were applied to improve the model's robustness and generalization capabilities.

## 3.4 Feature Extraction Using Vision Transformers

The feature extraction process leverages the Vision Transformer (ViT) architecture, which is pre-trained on the ImageNet dataset. The ViT model processes each input image by dividing it into patches and applying a transformer-based approach to analyze these patches, effectively capturing intricate visual patterns. To tailor the model for the specific task of plant species classification, the top layers of the pre-trained ViT model were removed. In their place, a custom classifier head was added, allowing the model to be fine-tuned and optimized for distinguishing between various plant species.

## 3.5 Transfer Learning and Fine-Tuning

The transfer learning process begins by freezing the initial layers of the pre-trained Vision Transformer model to retain its general feature extraction capabilities. This allows the model to leverage its existing knowledge while focusing on plant-specific features in the subsequent layers. The final layers are then fine-tuned using the combined plant leaf dataset, ensuring the model adapts to the nuances of the new data. This approach not only reduces training time but also significantly improves model accuracy, as it builds upon the pre-trained weights to achieve faster convergence and better performance with a relatively limited dataset.



## 3.6 Model Training

- **Classifier Setup:** A custom classifier was added to the Vision Transformer to adapt it for plant classification.  
*Linear layer (512 units) → ReLU activation → Dropout (0.5) → Output layer (matching number of classes).*
- **Optimizer:** Used *Adam optimizer* with a learning rate of 0.001.
- **Loss Function:** Applied *Cross-entropy loss* for multi-class classification.
- **Learning Rate Scheduler:** Decreased the learning rate by 0.1 every 5 epochs.
- **Early Stopping:** Implemented with a patience of 3 epochs to prevent overfitting.

This training strategy efficiently utilized the pre-trained model for accurate plant species classification, minimizing overfitting and optimizing convergence.

## 3.7 Evaluation Metrics

The performance of the fine-tuned model was evaluated using the following metrics:

- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The accuracy of positive predictions.
- **Recall:** The ability of the model to identify all relevant instances.
- **F1-Score:** A harmonic mean of precision and recall.
- A **confusion matrix** was used to visualize misclassifications across the 25 classes.

## 3.8 Plant Information Retrieval using Google Gen-AI

In addition to classification, the model integrates an AI chatbot using the Gemini Pro LLM model to provide users with detailed information on plant species. The chatbot interface allows users to interactively query information related to the identified plant species, such as ecological significance, medicinal uses, and cultivation details.

### 3.9 Computational Environment

The experiments were conducted using the **PyTorch** framework along with **Hugging Face Transformers** for implementing the Vision Transformer model. The computational environment was set up on a high-performance GPU server:

- **Python version:** 3.11
- **Libraries:** torch, transformers, scikit-learn, matplotlib, seaborn
- **Hardware:** NVIDIA A100 GPUs with CUDA support for accelerated training.

### 3.10 Vision Transformer (ViT) Architecture

The Vision Transformer (ViT) model used in this project is optimized for high-dimensional data like images:

- **Patch Embedding:** The image is divided into patches, each treated as a sequence input to the transformer.
- **Transformer Layers:** Consist of multi-head self-attention and feed-forward networks that capture complex patterns.
- **Classifier Head:** A custom classifier was added to adapt the model for multi-class plant species classification.

### 3.10 Vision Transformer (ViT) Architecture

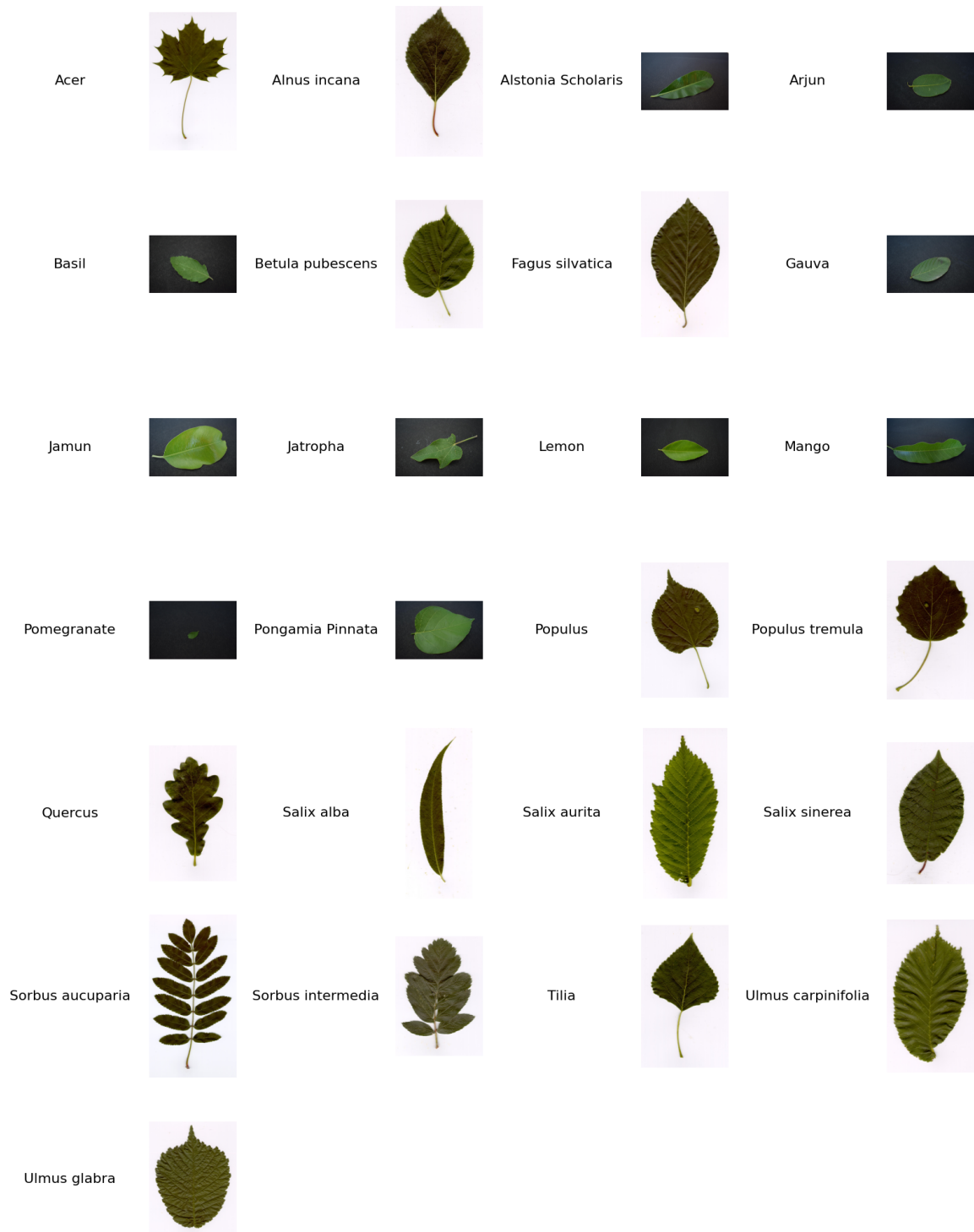


Fig. 3.1 Sample images from the expanded Swedish Leaf Dataset.[3]

### 3.10 Vision Transformer (ViT) Architecture

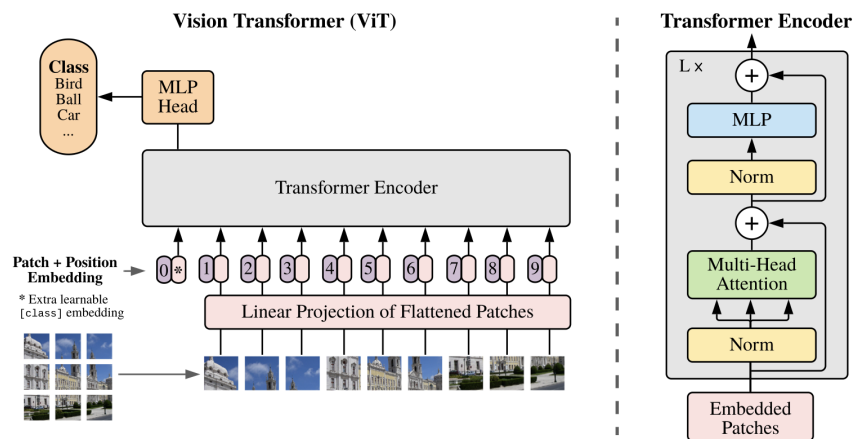


Fig. 3.2 Vision Transformer architecture showcasing patch embeddings and transformer layers. [1]

# Chapter 4

## Results

### 4.1 Confusion Matrix Result

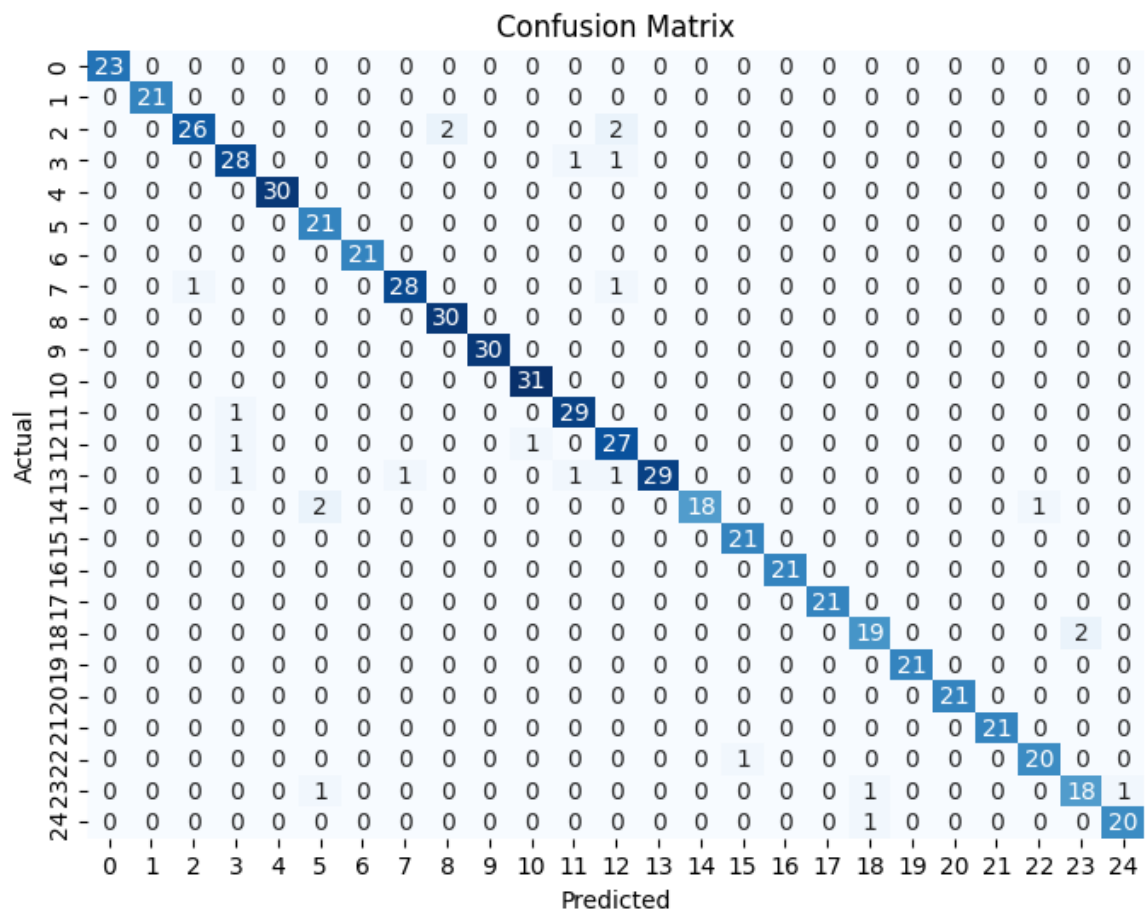



Fig. 4.1 Confusion Matrix

## 4.2 Model Training Result

```
Epoch 1/20, Training Loss: 2.1874
Validation Loss: 0.9478, Accuracy: 0.8452
Epoch 2/20, Training Loss: 0.8483
Validation Loss: 0.4636, Accuracy: 0.8677
Epoch 3/20, Training Loss: 0.5456
Validation Loss: 0.3156, Accuracy: 0.9403
Epoch 4/20, Training Loss: 0.4128
Validation Loss: 0.2514, Accuracy: 0.9484
Epoch 5/20, Training Loss: 0.3164
Validation Loss: 0.2217, Accuracy: 0.9403
Epoch 6/20, Training Loss: 0.2795
Validation Loss: 0.1996, Accuracy: 0.9597
Epoch 7/20, Training Loss: 0.2570
Validation Loss: 0.1911, Accuracy: 0.9532
Epoch 8/20, Training Loss: 0.2506
Validation Loss: 0.1873, Accuracy: 0.9581
Epoch 9/20, Training Loss: 0.2501
Validation Loss: 0.1845, Accuracy: 0.9597
Epoch 10/20, Training Loss: 0.2447
Validation Loss: 0.1826, Accuracy: 0.9581
Epoch 11/20, Training Loss: 0.2372
Validation Loss: 0.1783, Accuracy: 0.9597
Epoch 12/20, Training Loss: 0.2412
Validation Loss: 0.1798, Accuracy: 0.9597
Epoch 13/20, Training Loss: 0.2375
Validation Loss: 0.1775, Accuracy: 0.9597
Epoch 14/20, Training Loss: 0.2353
Validation Loss: 0.1796, Accuracy: 0.9597
Epoch 15/20, Training Loss: 0.2330
Validation Loss: 0.1806, Accuracy: 0.9597
Epoch 16/20, Training Loss: 0.2372
Validation Loss: 0.1768, Accuracy: 0.9597
Epoch 17/20, Training Loss: 0.2330
Validation Loss: 0.1780, Accuracy: 0.9597
Epoch 18/20, Training Loss: 0.2379
Validation Loss: 0.1785, Accuracy: 0.9597
Epoch 19/20, Training Loss: 0.2311
Validation Loss: 0.1768, Accuracy: 0.9597
Early stopping triggered.
```

Fig. 4.2 Transformer Model Training

## 4.3 Other Results



```
Precision: 0.96, Recall: 0.96, F1 Score: 0.96
```

Fig. 4.3 Other Result

## 4.4 Conclusion

This structured approach demonstrates the effectiveness of Vision Transformers and transfer learning in automating plant leaf classification, significantly reducing manual effort in species identification. Additionally, the integration of a chatbot powered by Google Generative AI (Gemini Pro) enhances user engagement by providing real-time, informative responses about identified plant species.

## 4.5 Future Work

- **Multilingual Support:** Implementing multiple language options in both the classification model and chatbot to broaden accessibility, especially in regions critical to plant conservation and research.
- **Disease Detection:** Expanding the model to identify plant diseases from leaf symptoms, enabling early intervention in agriculture and aiding conservation efforts.

# References

- [1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *Journal of Images*. viii, 4, 12
- [2] Gawli, K. S. and Gaikwad, A. S. (2020). Deep learning for plant species classification. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 7(11):99–104. ISSN-2349-5162. 5
- [3] Kaur, S. and Kaur, P. (2019). Plant species identification based on plant leaf using computer vision and machine learning techniques. *Journal of Multimedia Information System*, 6:49–60. viii, 4, 7, 11