LAB 1 Report

Course Name: Machine Learning

Course Code: CSE-475

Section - 3

Lab Name: Mango Leaf Diseases Classification using Decision Tree and Random Forest

Submitted By
Name: Ismail Mahmud Nur
ID: 2021-2-60-052
Dept. of Computer Science & Engineering

Submitted To
Dr Raihan Ul Islam,
Associate Professor,
Department of Computer Science and Engineering
East West University

Mango Leaf Disease Classification Using Machine Learning

Abstract

This lab assignment proposes a machine learning approach for the classification of mango leaf diseases using the Decision Tree and Random Forest models. Thus, in this way, we have analyzed the visual and color characteristics of images in different classes of disease via EDA. The dataset was balanced in classes, which allowed generalization of the model by training and testing. From the results obtained, it can be stated that the Random Forest model outperforms the Decision Tree model by 85%, which is very encouraging for its application in disease diagnosis.

1. Introduction

Mango is one of the most important fruit crops in tropical and subtropical regions, and it has been facing numerous challenges from leaf diseases that reduce yield and affect quality. Accurate diagnosis of these diseases is very essential for effective management. In this context, machine learning models—Decision Tree and Random Forest—are used for the classification of mango leaf diseases based on color and texture characteristics. Our goal here is to develop a model that can accurately distinguish between various disease categories and healthy leaves.

2. Methodology

2.1 Dataset Description

The dataset contains images of mango leaves from eight different classes: anthracnose, bacterial canker, cutting weevil, die back, gall midge, healthy, powdery mildew, and sooty mould. Each class contains roughly 500 photos. Preprocessing involves downsizing all pictures and extracting average RGB color intensities.

2.2 Exploratory Data Analysis (EDA)

The EDA phase involved assessing each category's distributional and visual features. The essential components were to present example photos from each category, analyze category distribution, visualize correlations between RGB channels, and calculate average RGB color intensities for each category.

The EDA phase involved:

- **Sample Image Display:** Displaying representative images from each category to understand visual differences.

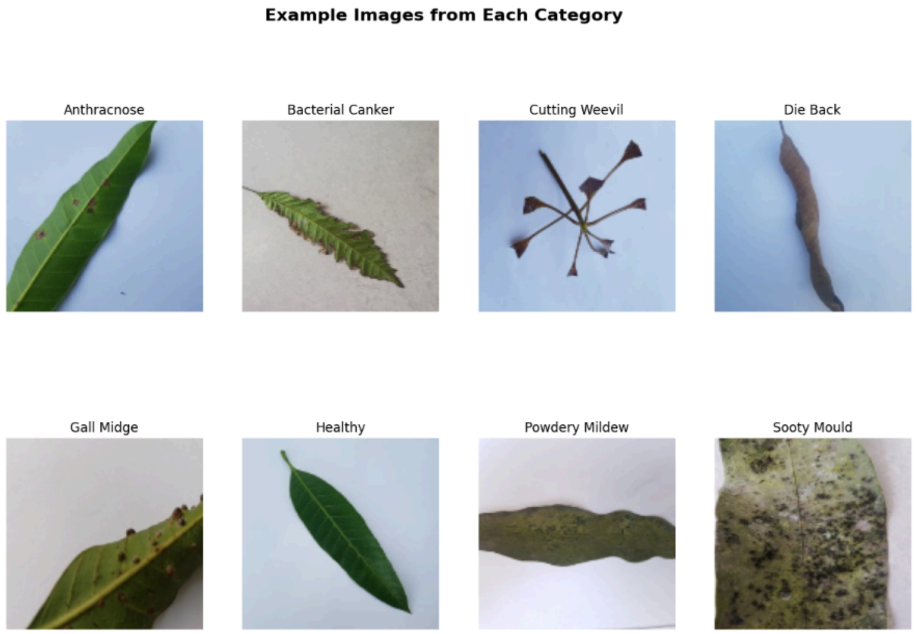


Figure 1: Example Image per Category

- **Category Distribution:** A balanced dataset with an equal number of images per category.

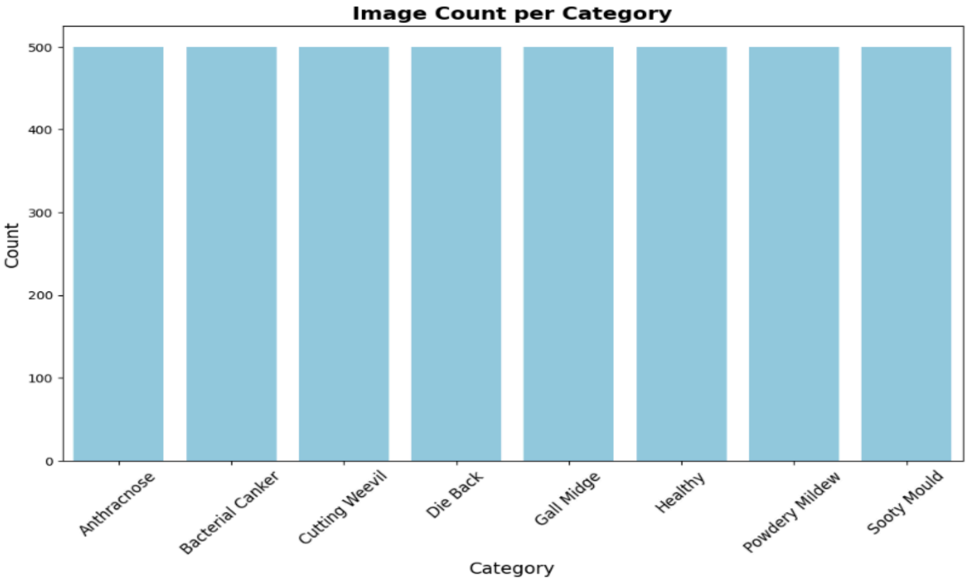


Figure 2: Image Count per Category

- **RGB Channel Correlation Heatmap:** Analyzing color intensity correlations across channels.

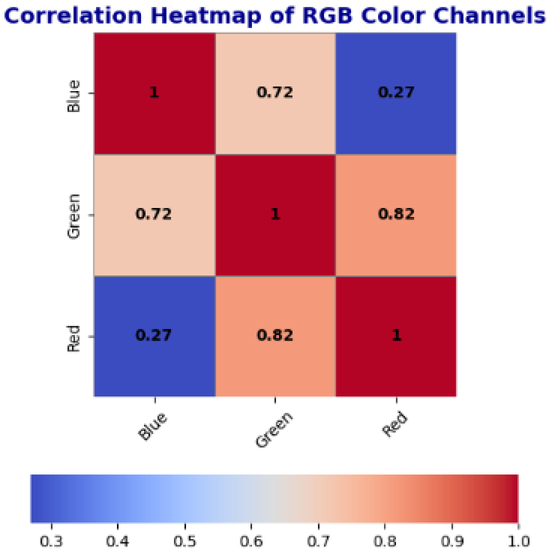
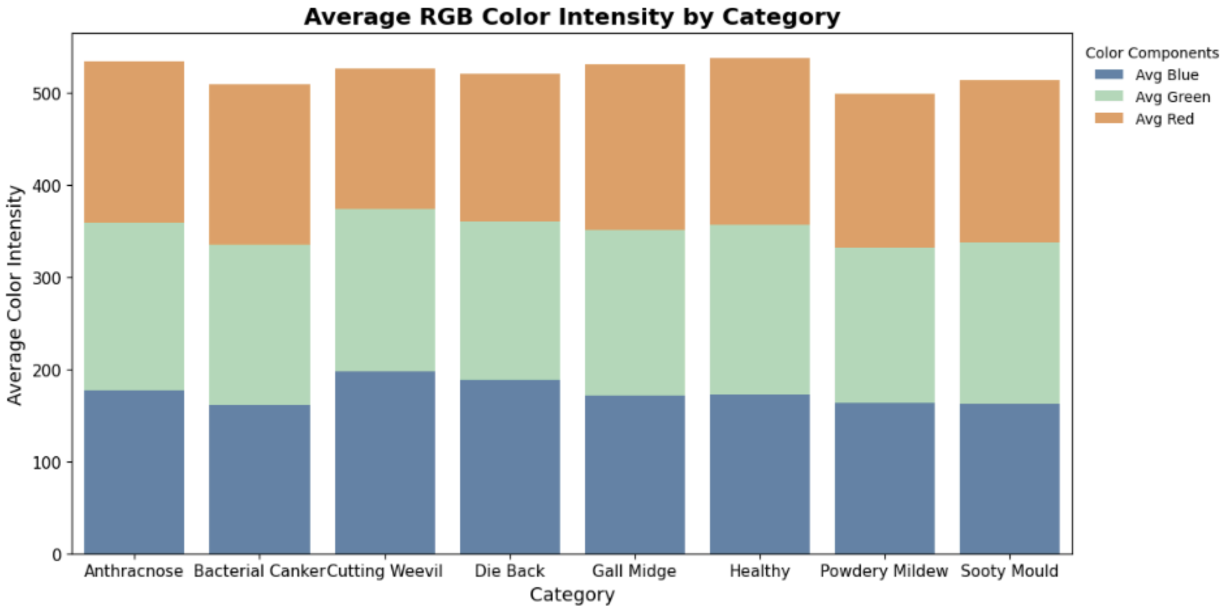


Figure 3: Correlation Heatmap of RGB Color Channels

- **Average RGB Intensities by Category:** Observing color intensity trends across categories.



The bars represent the average color intensity in each category's images for Blue, Green, and Red channels.
Higher values indicate a stronger presence of that color in the category's images.

Figure 4: Average RGB Color Intensity by Category

Category	Image Count	Avg Blue	Avg Green	Avg Red
Anthracnose	500	177.38	181.29	175.88
Bacterial Canker	500	161.36	173.47	174.16
Cutting Weevil	500	198.10	176.39	152.02
Die Back	500	188.66	171.64	160.61
Gall Midge	500	171.36	179.48	180.38
Healthy	500	173.12	184.13	182.07
Powdery Mildew	500	164.24	168.02	166.44
Sooty Mould	500	162.74	175.10	175.70

Table 1: Summary of Image Count and Average RGB Intensities by Category

3. Model Training and Evaluation

3.1 Model Descriptions

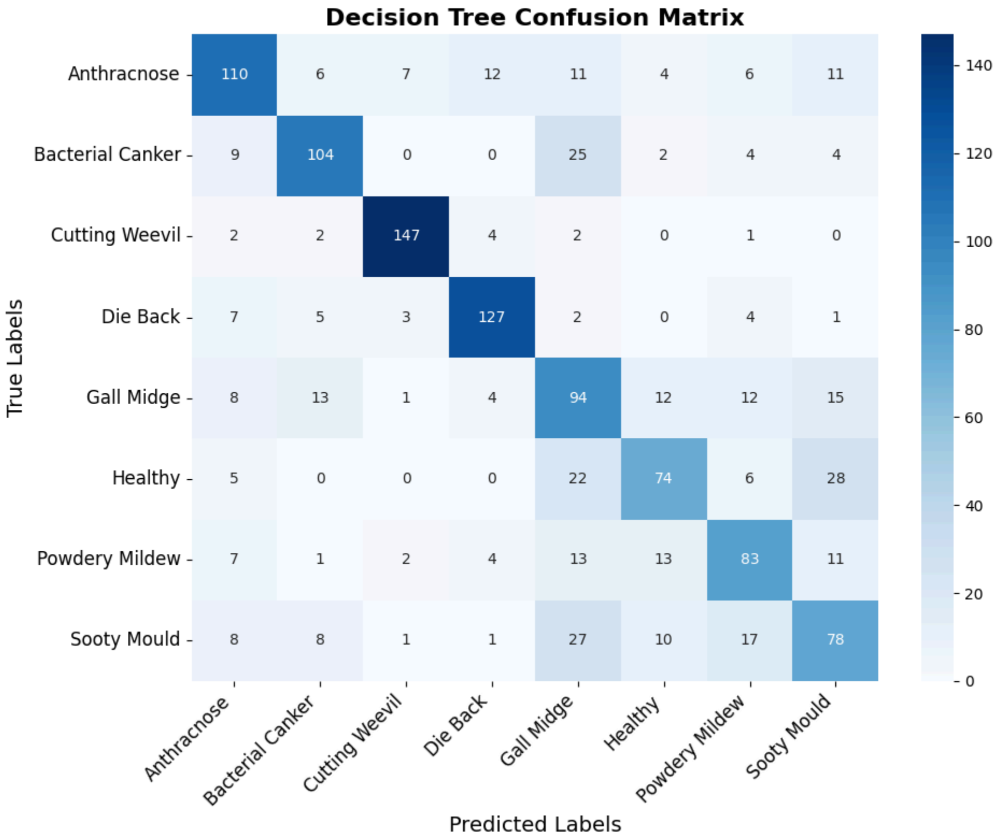
Two of the most popular classification models were selected: Decision Tree and Random Forest. Decision Tree is a single-tree classifier; it has interpretability but can suffer from overfitting. On the other hand, Random Forest is an ensemble of multiple decision trees, which usually has higher general accuracy and robustness against overfitting.

3.2 Model Training and Performance Metrics

The dataset was divided into a 70-30 ratio for training and testing, respectively. The performance metrics used to gauge this were: precision, recall, F1-score, and overall accuracy. The tables below show the actual results for each model: 2 and 3.

Category	precision	recall	f1-score	support
Anthracnose	0.71	0.66	0.68	167
Bacterial Canker	0.75	0.70	0.72	148
Cutting Weevil	0.91	0.93	0.92	158
Die Back	0.84	0.85	0.84	149
Gall Midge	0.48	0.59	0.53	159
Healthy	0.64	0.55	0.59	135
Powdery Mildew	0.62	0.62	0.62	134
Sooty Mould	0.52	0.52	0.52	150
Overall Accuracy : 0.68				

Table 2: Decision Tree Performance Metrics



	Anthracnose	Bacterial Canker	Cutting Weevil	Die Back	Gall Midge	Healthy	Powdery Mildew	Sooty Mould
Anthracnose	110	6	7	12	11	4	6	11
Bacterial Canker	9	104	0	0	25	2	4	4
Cutting Weevil	2	2	147	4	2	0	1	0
Die Back	7	5	3	127	2	0	4	1
Gall Midge	8	13	1	4	94	12	12	15
Healthy	5	0	0	0	22	74	6	28
Powdery Mildew	7	1	2	4	13	13	83	11
Sooty Mould	8	8	1	1	27	10	17	78

Figure 5: Decision Tree Confusion Matrix

Category	**precision**	**recall**	**f1-score**	**support**
Anthracnose	0.85	0.86	0.85	167
Bacterial Canker	0.82	0.78	0.80	148
Cutting Weevil	0.97	0.96	0.97	158
Die Back	0.91	0.95	0.93	149
Gall Midge	0.70	0.76	0.73	159
Healthy	0.90	0.89	0.90	135
Powdery Mildew	0.86	0.84	0.85	134
Sooty Mould	0.79	0.75	0.77	150
Overall Accuracy : 0.85				

Table 3: Random Forest Performance Metrics

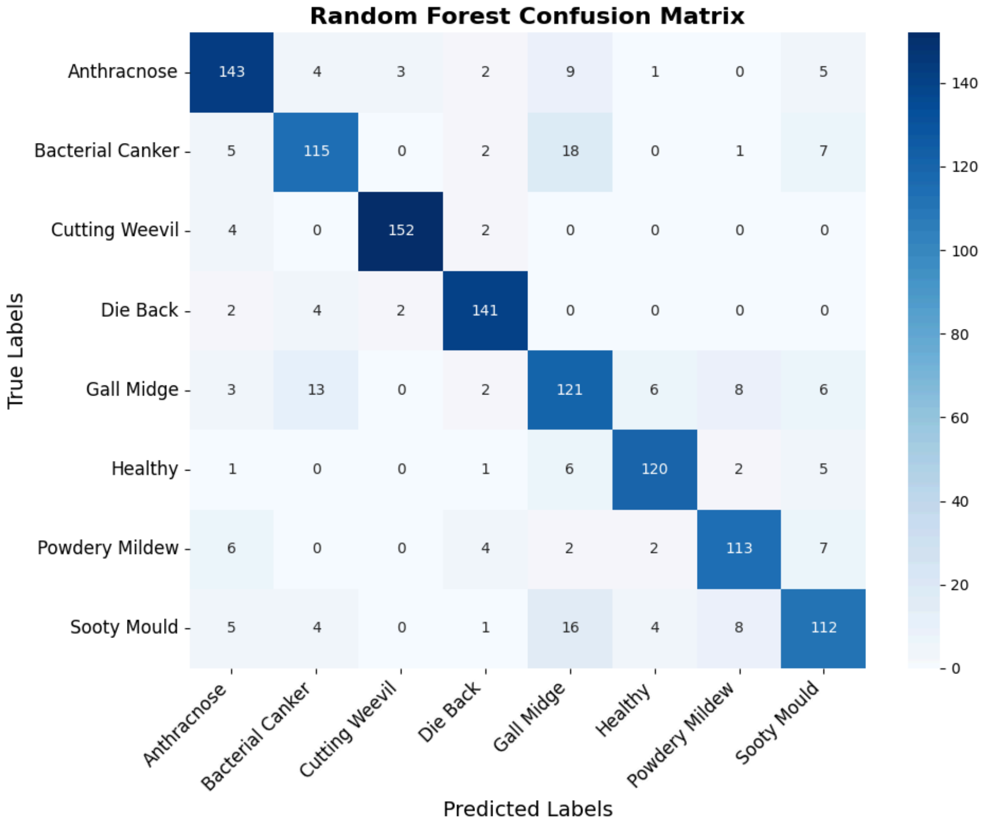


Figure 6: Random Forest Confusion Matrix

Summary Table of Model	Overall Accuracy
Decision Tree	0.68
Random Forest	0.85

Table 4: Model Accuracy Comparison Summary

4. Results

The Random Forest model achieved much better accuracy (85%) compared to the Decision Tree model (68%). The detailed metrics of performances show that, concerning precision, recall, and F1-score, Random Forest performs better than Decision Tree in almost all classes of diseases found on mango leaves. Comparing both models, it would seem that the Random Forest is more appropriate in disease classification of mango leaves, most likely due to its ensemble nature, reducing overfitting, and enabling the capture of more complex patterns.

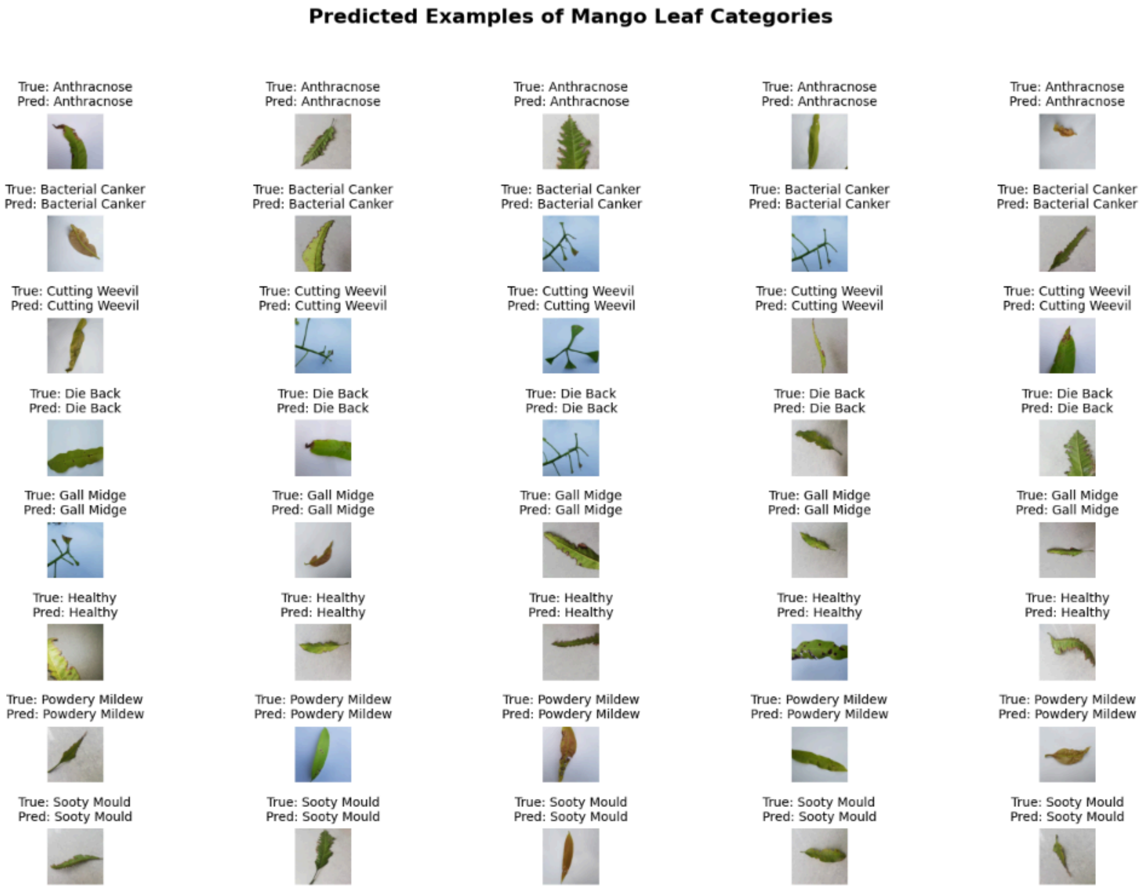

Predicted Examples of Mango Leaf Categories
True: Anthracnose Pred: Anthracnose
True: Anthracnose Pred: Anthracnose
True: Anthracnose Pred: Anthracnose
True: Anthracnose Pred: Anthracnose
True: Anthracnose Pred: Anthracnose
True: Bacterial Canker Pred: Bacterial Canker
True: Bacterial Canker Pred: Bacterial Canker
True: Bacterial Canker Pred: Bacterial Canker
True: Bacterial Canker Pred: Bacterial Canker
True: Bacterial Canker Pred: Bacterial Canker
True: Cutting Weevil Pred: Cutting Weevil
True: Cutting Weevil Pred: Cutting Weevil
True: Cutting Weevil Pred: Cutting Weevil
True: Cutting Weevil Pred: Cutting Weevil
True: Cutting Weevil Pred: Cutting Weevil
True: Die Back Pred: Die Back
True: Die Back Pred: Die Back
True: Die Back Pred: Die Back
True: Die Back Pred: Die Back
True: Die Back Pred: Die Back
True: Gall Midge Pred: Gall Midge
True: Gall Midge Pred: Gall Midge
True: Gall Midge Pred: Gall Midge
True: Gall Midge Pred: Gall Midge
True: Gall Midge Pred: Gall Midge
True: Healthy Pred: Healthy
True: Healthy Pred: Healthy
True: Healthy Pred: Healthy
True: Healthy Pred: Healthy
True: Healthy Pred: Healthy
True: Powdery Mildew Pred: Powdery Mildew
True: Powdery Mildew Pred: Powdery Mildew
True: Powdery Mildew Pred: Powdery Mildew
True: Powdery Mildew Pred: Powdery Mildew
True: Powdery Mildew Pred: Powdery Mildew
True: Sooty Mould Pred: Sooty Mould
True: Sooty Mould Pred: Sooty Mould
True: Sooty Mould Pred: Sooty Mould
True: Sooty Mould Pred: Sooty Mould
True: Sooty Mould Pred: Sooty Mould

Figure 7: Predicted Examples of Mango Leaf Categories

5. Discussion

Our analysis has demonstrated the ability of color intensity and simple machine learning models in classifying mango leaf diseases, where the superior performance of the Random Forest can be traced back to its generalization capability over a single-tree classifier. The EDA also shows that some of the disease categories have specific patterns of colors, which help in the classification by the models. More features like texture or color histograms can be tried in future works, and some deep learning methods may be followed for better accuracy.

6. Conclusion

The lab assignment indeed managed to classify diseases of the mango leaf using machine learning models, where Random Forest gave the best accuracy. The model can most probably help with real-world disease detection and early intervention for proper crop management. Another alternative is to study more complex models and other visual aspects to further improve the findings.