DataScientest

Data science-Project: Plant Recognition

Comprehensive Analysis of Plant Recognition Datasets

Project mentor:

Roman Lesieur

Project participants:

Fares Naem

Mohammadamin Nooralikheirzad

Nathalie Zahran

Philippe Maitey Masindet

20.06.2024

TABLE OF CONTENTS

1. Introduction

In modern agriculture, the detection and treatment of plant diseases are critical for ensuring healthy crop yields and minimizing wasteful practices. Traditional methods, such as overhead pesticide spraying using helicopters or planes, often result in excessive use of chemicals, causing environmental harm and economic inefficiency. This project aims to revolutionize plant disease management by developing an AI-based system which can be integrated into robotic platforms to detect and treat diseased plants with precision.

The primary goal of this project is to create a robust computer vision model capable of accurately identifying various plant diseases from images in real-time. By utilizing advanced image processing techniques, the system will enhance disease detection accuracy, even under varying lighting and environmental conditions. This model can then be integrated into a robotic platform.

To ensure the system's scalability and adaptability, the AI model will be designed to accommodate different crops and diseases. This adaptability will facilitate the widespread adoption of precision agriculture practices, ultimately contributing to sustainable farming and enhanced food security.

By integrating our AI -model in agriculture, this project aims to transform plant disease management, making it more efficient, precise, and environmentally friendly. Through the successful implementation of this technology, farmers can achieve healthier crops, reduce chemical usage, and improve agricultural productivity.

This report provides a comprehensive analysis of various plant recognition datasets, highlighting their characteristics, the diversity of plant types, image qualities, and model performance.

The datasets included are:

- **New Plant Disease Dataset**
- **Plant Disease Dataset**
- **Plant Village Dataset**

The datasets excluded is:

- **V2 Plant Seedlings Dataset**

This dataset lacks information about the health status of the plants (e.g., healthy, specific disease types), which is essential for our project aimed at classifying both plants and their diseases.

The V2 Plant Seedlings dataset has more variability in image dimensions, which may require resizing for uniformity in model training unlike the other datasets that have a consistent size of 256x256 pixels, especially in the New Plant Disease, Plant Disease, and Plant Village datasets

Moreover, it contains seedling plant types that are not present in our other datasets. Consequently, we cannot use this dataset to train our Convolutional Neural Network (CNN) model for the project. However, there will be a few graphical depictions of the V2 Plant Seedlings Dataset, to better ground our reasoning.

2. Datasets Overview

The datasets analyzed in this report are sourced from Kaggle:

- **V2 Plant Seedlings Dataset**: Contains images of different plant seedlings.

  https://www.kaggle.com/datasets/vbookshelf/v2-plant-seedlings-dataset

- **New Plant Diseases Dataset**: Contains images of plant diseases.

  https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset

- **Plant Disease Dataset**: Like the New Plant Disease dataset but with fewer images.

  https://www.kaggle.com/datasets/saroz014/plant-disease

- **Plant Village Dataset**: Contains images in three formats: color, grayscale, and segmented.

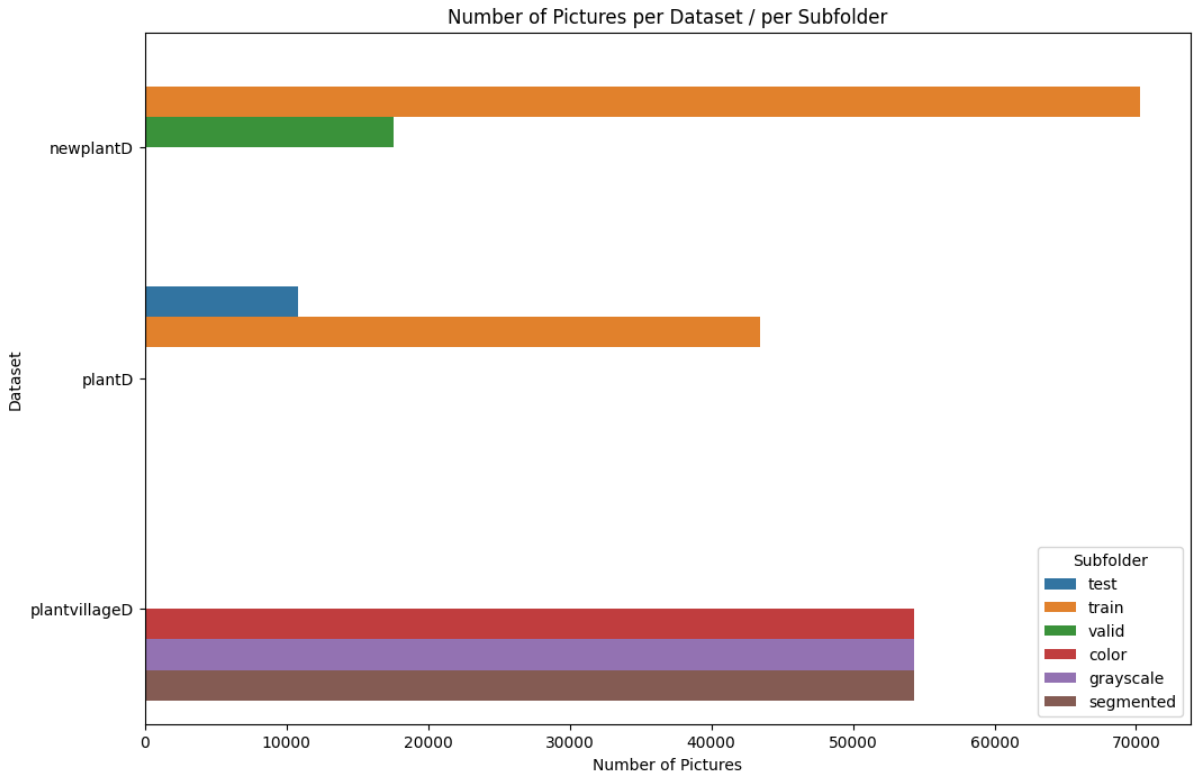
  https://www.kaggle.com/datasets/abdallahalidev/plantvillage-dataset

These datasets vary significantly in the number of images and the types of plants they cover. The V2 Plant Seedlings dataset focuses on seedling stages, while the other datasets cover various stages of plant diseases.

3. File Size and Image Dimensions

The file size and image dimensions are important factors that affect the processing and analysis of images in machine learning models.

File Size Distribution: The file sizes vary across datasets, affecting storage and computational requirements. Bar plots can effectively visualize the distribution of file sizes across datasets.



Below is a short description of the datasets.

I) Plant Disease:

Consists of a test subfolder with 10,850 images, and a training subfolder with 43,457 images.

Observation: The training set significantly outweighs the test set, common in datasets aimed at building robust predictive models.

II) Plant Village:

This dataset is divided into three subfolders, each dedicated to different image modifications such as, in color, in grayscale and segmented, with each containing around 54,307 images.

Observations: uniform distribution across different processing types suggests a focus on versatility in model training to accommodate various imaging conditions.

III) New Plant Diseases

It features a smaller test set with 33 images, a large training set with 70,297 images, and a validation set containing 17,573 images

Observations: The unusual large validation set relative to the test set indicates a focus on extensive model validation, possibly due to the complex variety of the diseases, and the conditions included.

Image Dimensions: Most images in the datasets have a consistent size of 256x256 pixels, especially in the New Plant Disease, Plant Disease, and Plant Village datasets.

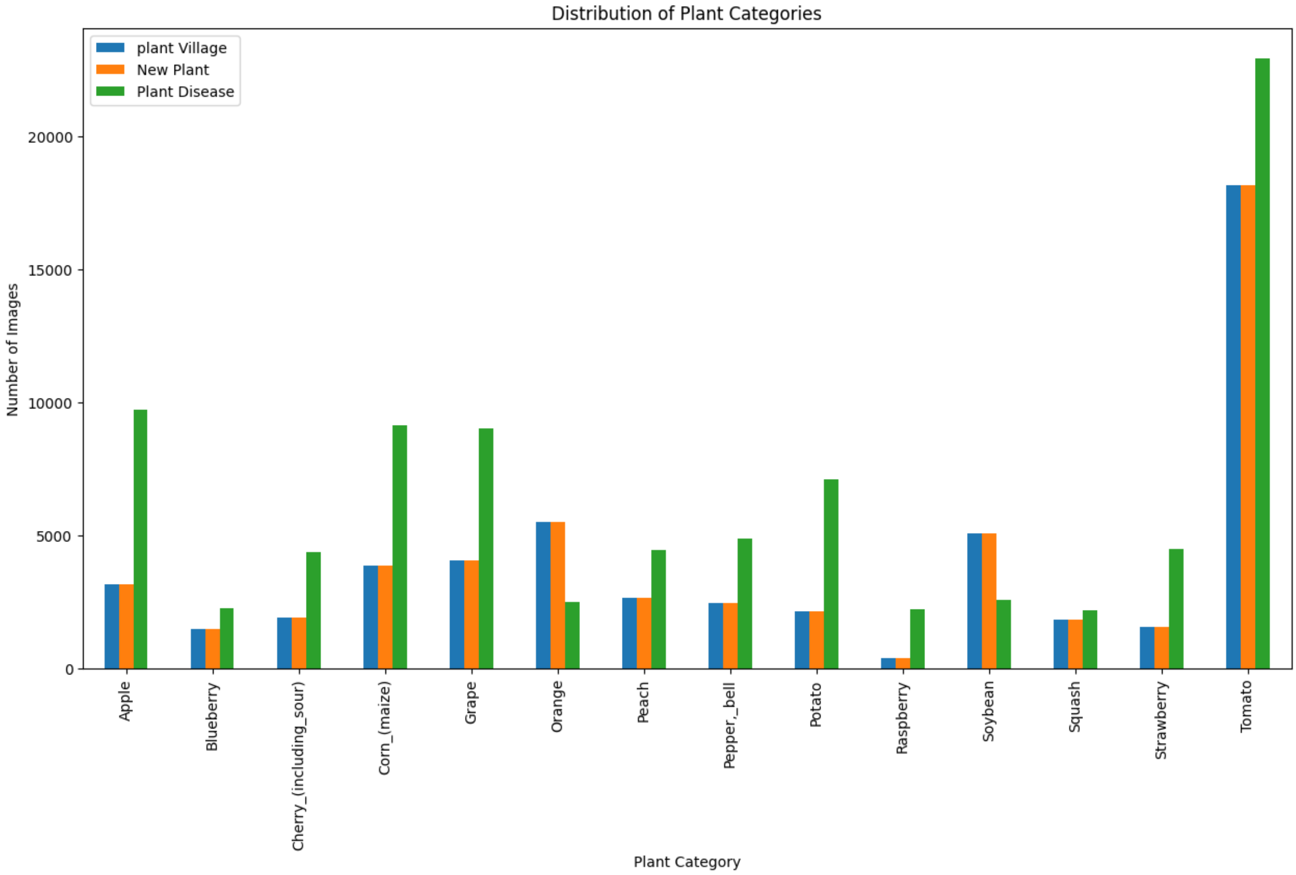
4. Plant Types Distribution

Understanding the distribution of plant types within each dataset is crucial for ensuring balanced model training and evaluation.

Dataset Specifics:

- **New Plant Disease and Plant Disease**: Cover the same plant types but differ in the number of images.
- **Plant Village**: Contains the same plant types as the disease datasets but includes segmented images, images in color and in the grayscale, this provides additional insights into plant features.

The graph displays the image distribution for various plant categories across three datasets: Plant Village, New Plant Diseases, and Plant Disease.



The Critical observations include:

- **Imbalance in Representation:** The abundance of tomato images, particularly in the New Plant and Plant Disease datasets, may bias model training towards tomato-specific occurences, potentially reducing the model's effectiveness on other plants.
- **Underrepresented Categories:** Crops such as squash and raspberry are underrepresented across the datasets. This insufficiency of data could lead to

underperforming models for these categories, impacting the models' utility in diverse agricultural settings.

- **Variability in Crop Focus:** Corn and Apple are well-represented in the New Plant dataset but not as much in Plant Village and Plant Disease. This variability suggests that the choice of dataset could significantly impact model performance depending on the crop of interest.

5. Distribution of Healthy and Diseased Plant Images

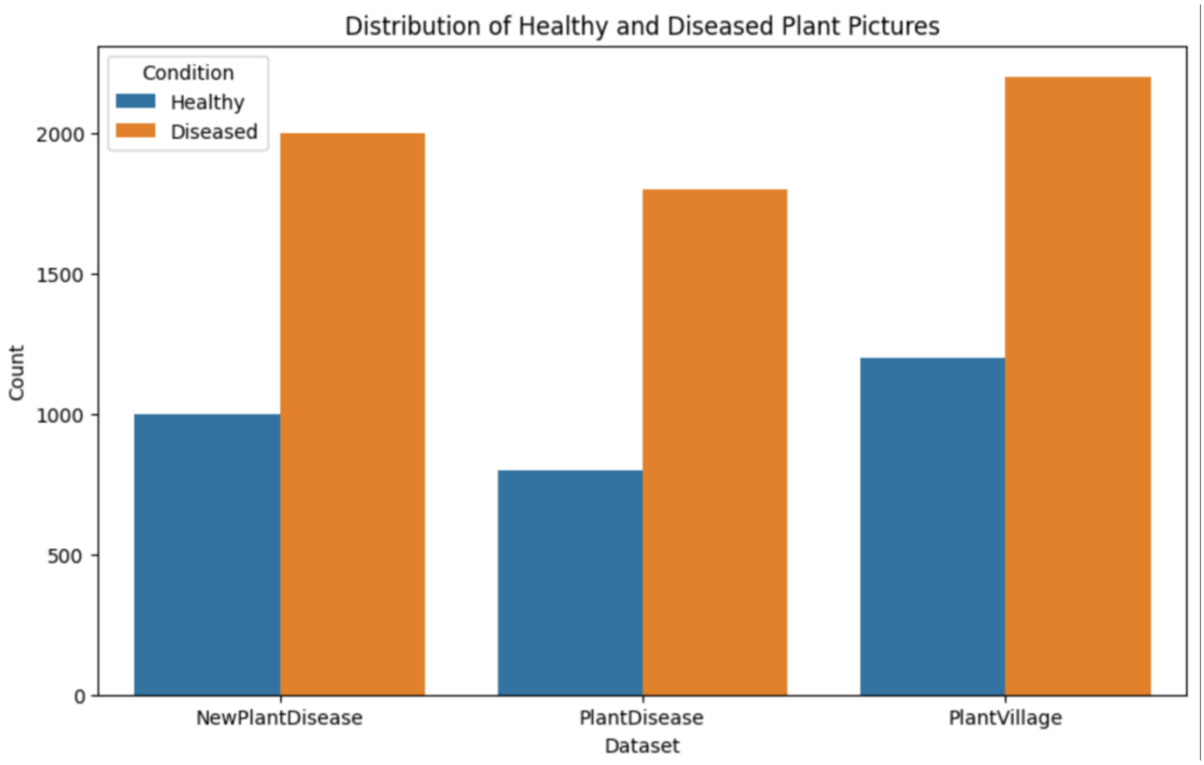
The distribution of healthy and diseased plant pictures varies across datasets. The New Plant Disease, Plant Disease, and Plant Village datasets have a higher proportion of diseased images compared to healthy ones. This imbalance needs to be addressed during model training to avoid bias.

Observation:

- **Healthy Only**: examples; Cherry healthy, Corn(maize) healthy etc.
- **Diseased Only**: examples; Apple scab, Grape black rot. etc.

This indicates that not all plant types have both healthy and diseased images, which may impact the model's ability to generalize across different conditions.

The images in the datasets can be simply categorized as either image depictions of healthy or diseased plants. The graph below visualizes the inequal distribution of the image types in the three used data sets.



We can therefore see that we would need to extract more of the plant images of the diseased type during the training and testing of our machine learning models.

Health Status Distribution by Plant Category

I) Imbalances in Health Status:

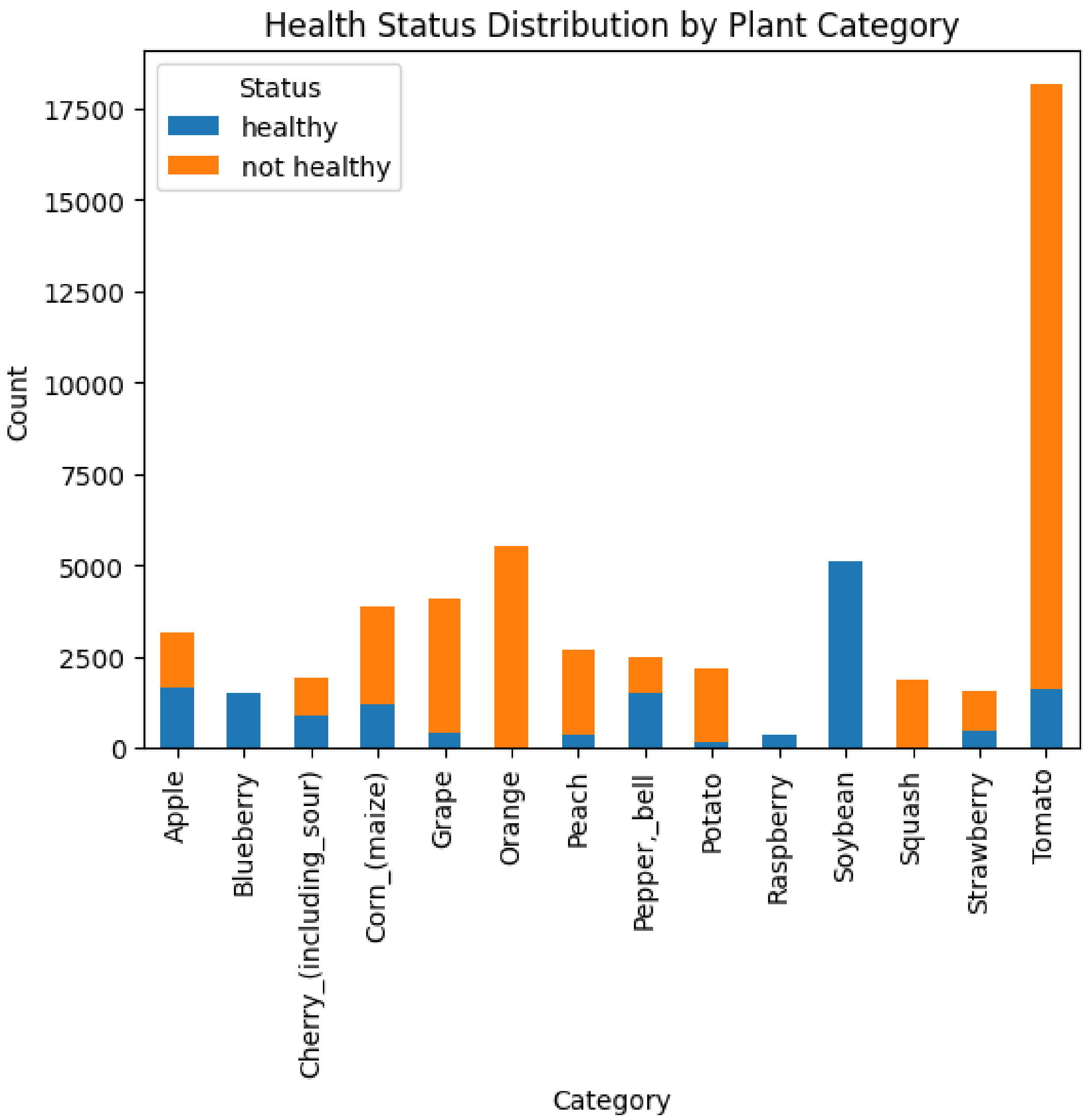
Tomato Dominance: The tomato category overwhelmingly contains not healthy samples, with almost no healthy samples represented. This heavy skew towards diseased samples may bias models towards disease detection, potentially reducing their effectiveness in identifying healthy tomato plants.
Neglected Healthy Samples: Several plant categories such as Apple, Blueberry, and Corn have significantly fewer healthy samples compared to not healthy ones. This distribution could lead to models that are overly sensitive to diseases and may falsely identify healthy plants as diseased.

II) Gaps in Dataset Coverage

Limited Data for Certain Crops: Crops like Squash and Raspberry show a drastic underrepresentation in both healthy and not healthy statuses. Such limited data points can lead to poor model performance for these categories, as the models have insufficient examples to learn from.
Bias Towards Common Crops: There is a noticeable focus on common crops such as Tomato and Corn, which may lead to models that perform well for these but poorly for less common crops due to the lack of balanced data.

III) Potential Model Limitations

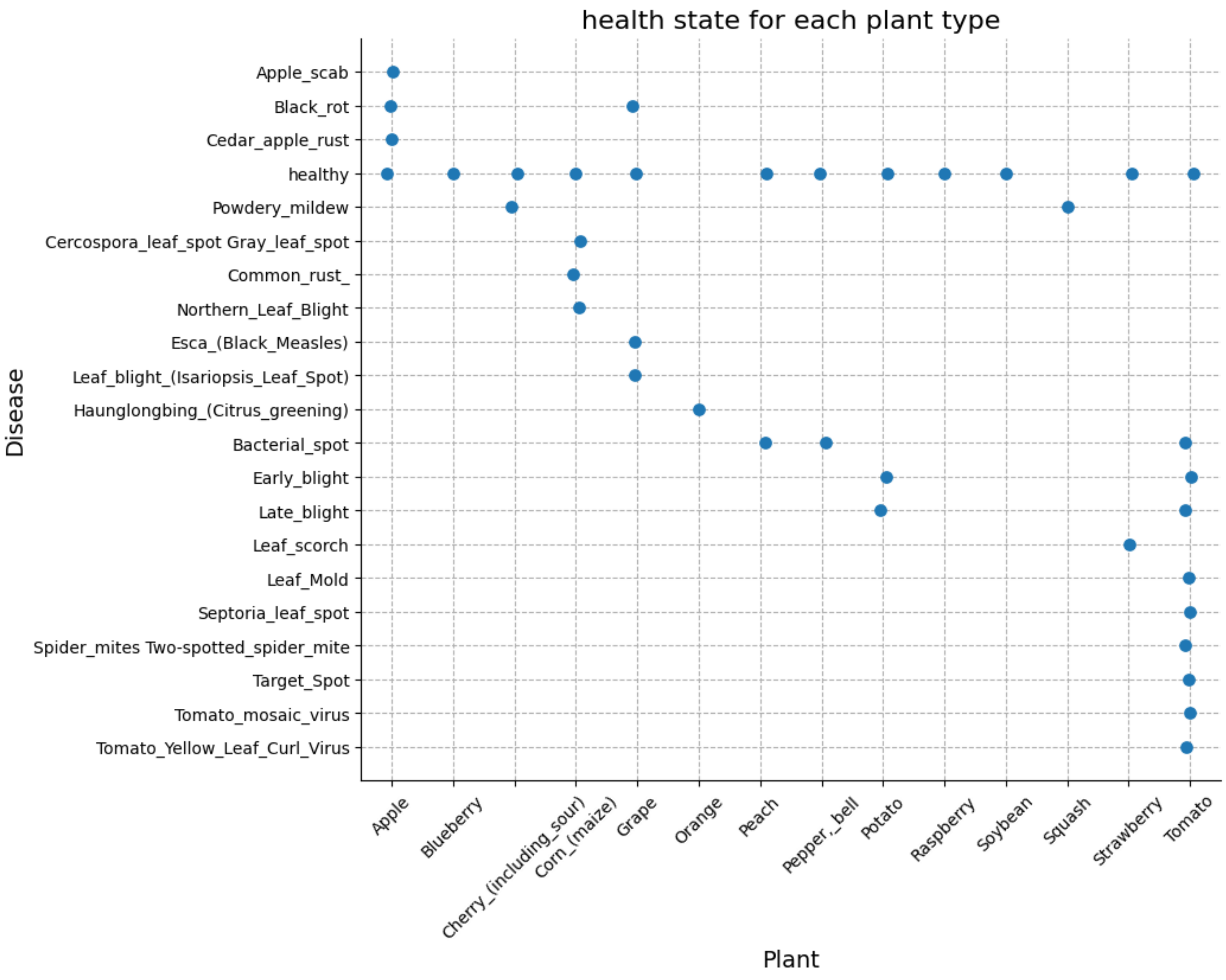
Overfitting Risk: With such a high representation of diseased samples, particularly in economically significant crops like Tomato, there is a risk that models trained on this data may overfit to specific diseases or symptoms seen in the dataset, limiting their generalizability to other conditions or new environments.
Underperformance on Healthy Plants: The scarcity of healthy plant data across most categories suggests that models may underperform in scenarios where distinguishing between genuinely healthy and subtly symptomatic plants is crucial.



Additionally, the following graph presents the health status of various plant types, indicating whether they are healthy or affected by a specific disease. Several interesting insights can be drawn from this graph. For instance, while most plant types have images representing both healthy and diseased states, some, like Orange and Squash, only have images for specific diseases (Bacterial spot and Powdery mildew, respectively).

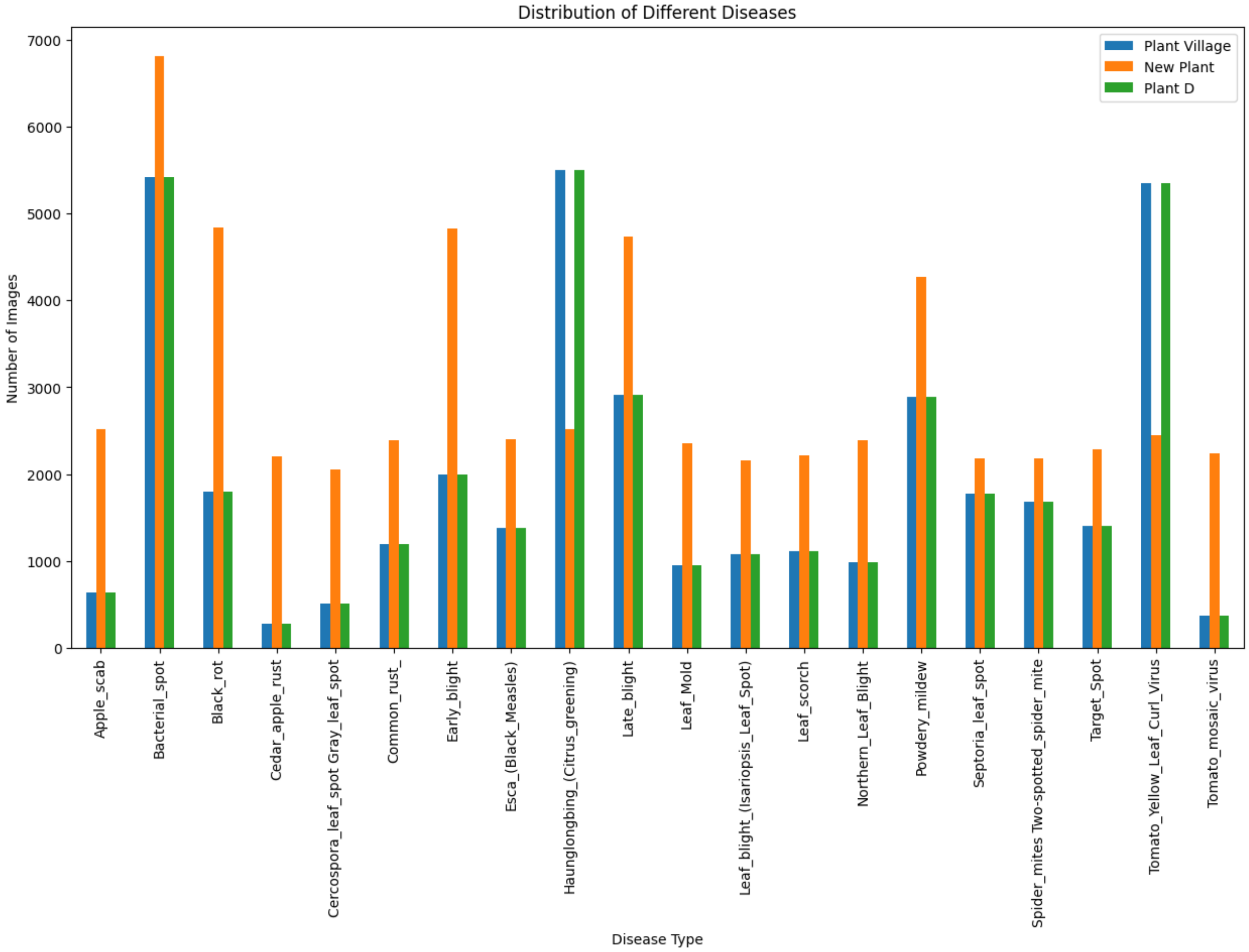
The graph also highlights common diseases across different plant types. For example, Black rot affects both Apple and Grape, Powdery mildew is seen in both Cherry and Squash, and Late blight is common to both Potato and Tomato. Additionally, it is evident from the graph how many disease types are recorded for each plant type in our dataset. Tomato, for instance, shows the highest number of disease types, whereas Grape is associated with only two diseases (Leaf blight and Huanglongbing).

This graph has been plotted using data from three datasets, and the observed patterns are consistent across all three.


health state for each plant type
Disease
Plant
Apple_scab
Black_rot
Cedar_apple_rust
healthy
Powdery_mildew
Cercospora_leaf_spot Gray_leaf_spot
Common_rust_
Northern_Leaf_Blight
Esca_(Black_Measles)
Leaf_blight_(Isariopsis_Leaf_Spot)
Haunglongbing_(Citrus_greening)
Bacterial_spot
Early_blight
Late_blight
Leaf_scorch
Leaf_Mold
Septoria_leaf_spot
Spider_mites Two-spotted_spider_mite
Target_Spot
Tomato_mosaic_virus
Tomato_Yellow_Leaf_Curl_Virus
Apple
Blueberry
Cherry_(including_sour)
Corn_(maize)
Grape
Orange
Peach
Pepper,_bell
Potato
Raspberry
Soybean
Squash
Strawberry
Tomato

6. Distribution of Different Diseases

The bar graph outlines the distribution of images for different diseases within the datasets



The key critical points are:

- **Disease-Specific Prevalence:** Certain diseases, like Apple Scab and Cedar Apple Rust, are overrepresented, particularly in the New Plant dataset. This can lead to models that are overly tuned to specific diseases, potentially at the expense of generalizability to less common diseases.
- **General Disease Coverage:** While some diseases show a balanced distribution, others, like Powdery Mildew and Leaf Mold, do not. This uneven distribution could lead to biased models that perform well on well-represented diseases but poorly on others.
- **Lack of Diversity in Disease Types:** The near-total uniformity of certain diseases within specific datasets may hinder the development of the models' capability of recognizing a broader spectrum of plant health conditions.

7. Color Distribution and Segmentation

Color distribution analysis helps in understanding the influence of background and segmentation on the images. The Plant Village dataset, with its segmented images, allows for a more focused analysis of the plant features without the background noise.

Healthy vs. Diseased Leaves:

Healthy Leaves: Typically show a balanced color distribution.

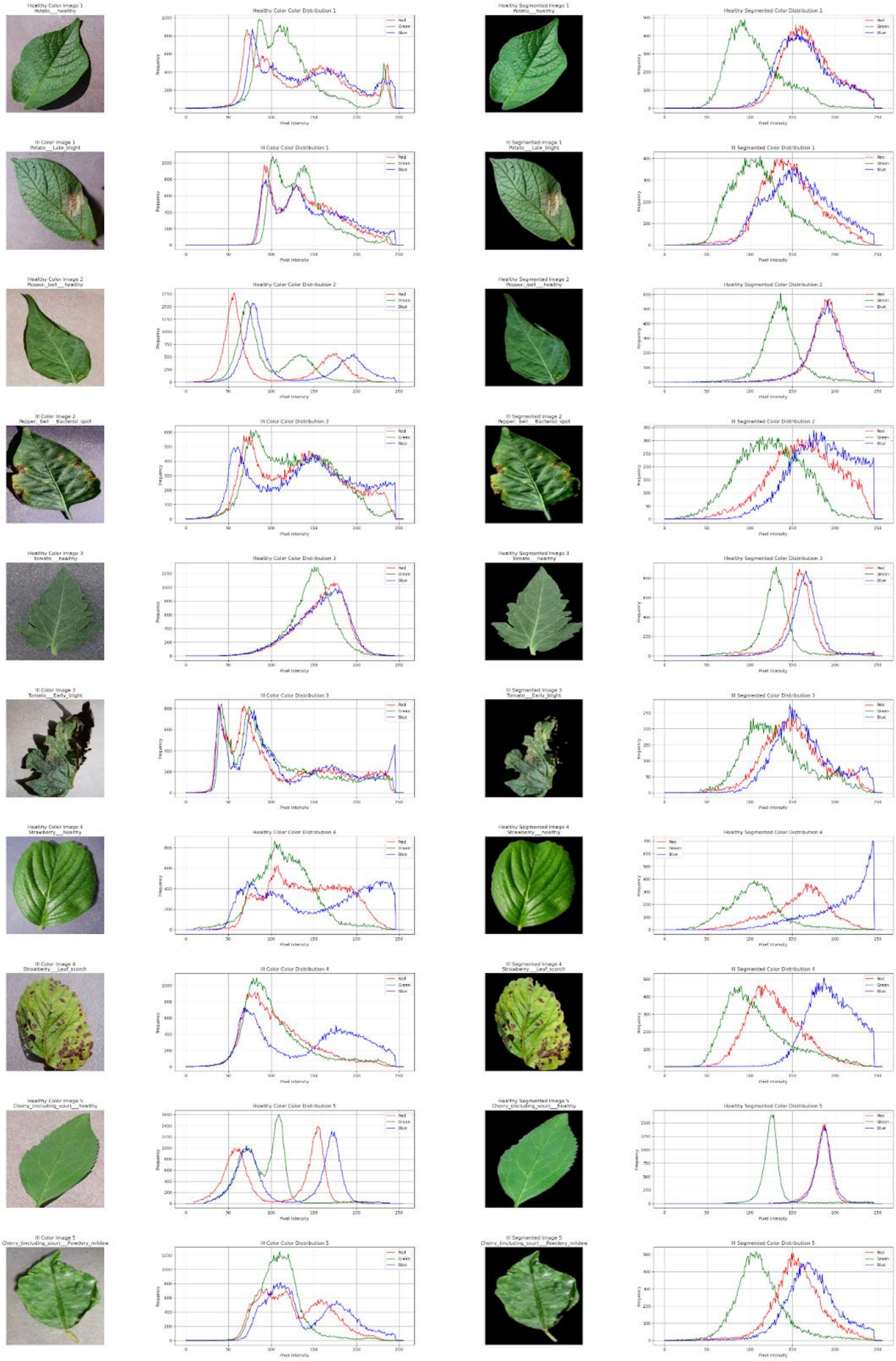
Diseased Leaves: May show imbalances, particularly in the red and blue channels.

The provided image underscores the importance of background segmentation in colour distribution analysis. Background elements in leaf images can significantly distort the colour distribution data, complicating the accurate identification and classification of plant diseases. By isolating the leaf from its background, the segmented images offer a clearer and more precise colour distribution histogram focusing exclusively on the leaf's characteristics. This enhanced clarity helps improve the accuracy of machine learning models used for disease detection by supplying cleaner, more relevant data inputs.

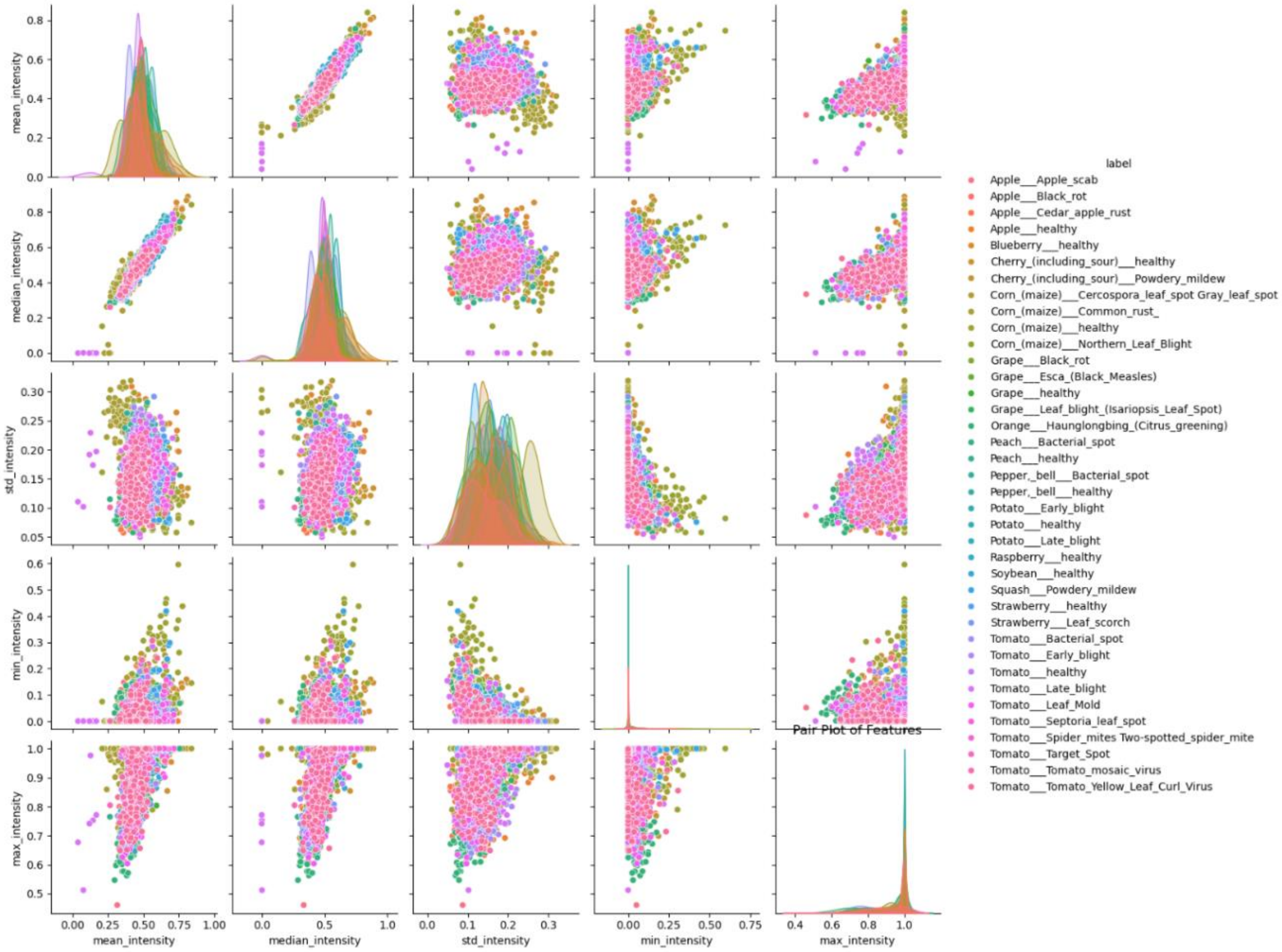
This image's empirical benefits are evident in its demonstration of how pre-processing steps, such as segmentation, can significantly enhance data quality.

The colour distribution histograms for the segmented images reveal marked differences in distribution curves, accurately reflecting the leaf's colour characteristics without background interference. This allows for a more precise identification of subtle differences between healthy and diseased leaves, leading to more accurate and reliable disease classification. Such visualizations are crucial for validating the pre-processing steps and guiding further refinements in the data processing pipeline.

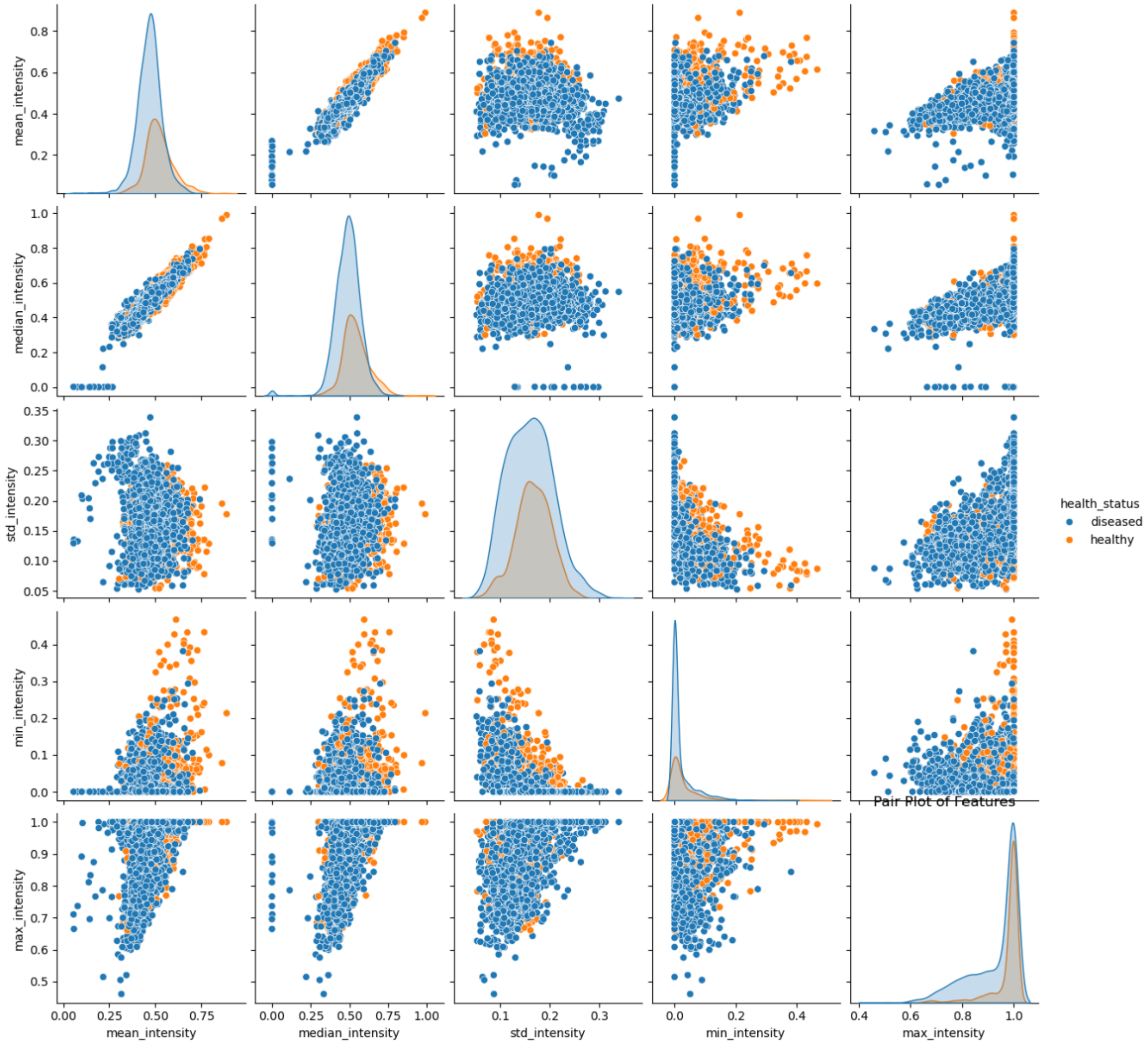
Including these segmented and non-segmented images in the report provides a practical example of how segmentation can enhance the accuracy of our plant disease classification system. This clear representation helps in conveying the importance of pre-processing to stakeholders, ensuring that the methodology and its benefits are well understood.


Healthy Color Image 1
Potato___healthy
Healthy Color Color Distribution 1
Healthy Segmented Image 1
Potato___healthy
Healthy Segmented Color Distribution 1
Ill Color Image 1
Potato___Late_blight
Ill Color Color Distribution 1
Ill Segmented Image 1
Potato___Late_blight
Ill Segmented Color Distribution 1
Healthy Color Image 2
Pepper,_bell___healthy
Healthy Color Color Distribution 2
Healthy Segmented Image 2
Pepper,_bell___healthy
Healthy Segmented Color Distribution 2
Ill Color Image 2
Pepper,_bell___Bacterial_spot
Ill Color Color Distribution 2
Ill Segmented Image 2
Pepper,_bell___Bacterial_spot
Ill Segmented Color Distribution 2
Healthy Color Image 3
Tomato___healthy
Healthy Color Color Distribution 3
Healthy Segmented Image 3
Tomato___healthy
Healthy Segmented Color Distribution 3
Ill Color Image 3
Tomato___Early_blight
Ill Color Color Distribution 3
Ill Segmented Image 3
Tomato___Early_blight
Ill Segmented Color Distribution 3
Healthy Color Image 4
Strawberry___healthy
Healthy Color Color Distribution 4
Healthy Segmented Image 4
Strawberry___healthy
Healthy Segmented Color Distribution 4
Ill Color Image 4
Strawberry___Leaf_scorch
Ill Color Color Distribution 4
Ill Segmented Image 4
Strawberry___Leaf_scorch
Ill Segmented Color Distribution 4
Healthy Color Image 5
Cherry_(including_sour)___healthy
Healthy Color Color Distribution 5
Healthy Segmented Image 5
Cherry_(including_sour)___healthy
Healthy Segmented Color Distribution 5
Ill Color Image 5
Cherry_(including_sour)___Powdery_mildew
Ill Color Color Distribution 5
Ill Segmented Image 5
Cherry_(including_sour)___Powdery_mildew
Ill Segmented Color Distribution 5
Red
Green
Blue
Frequency
Pixel Intensity
0
50
100
150
200
250

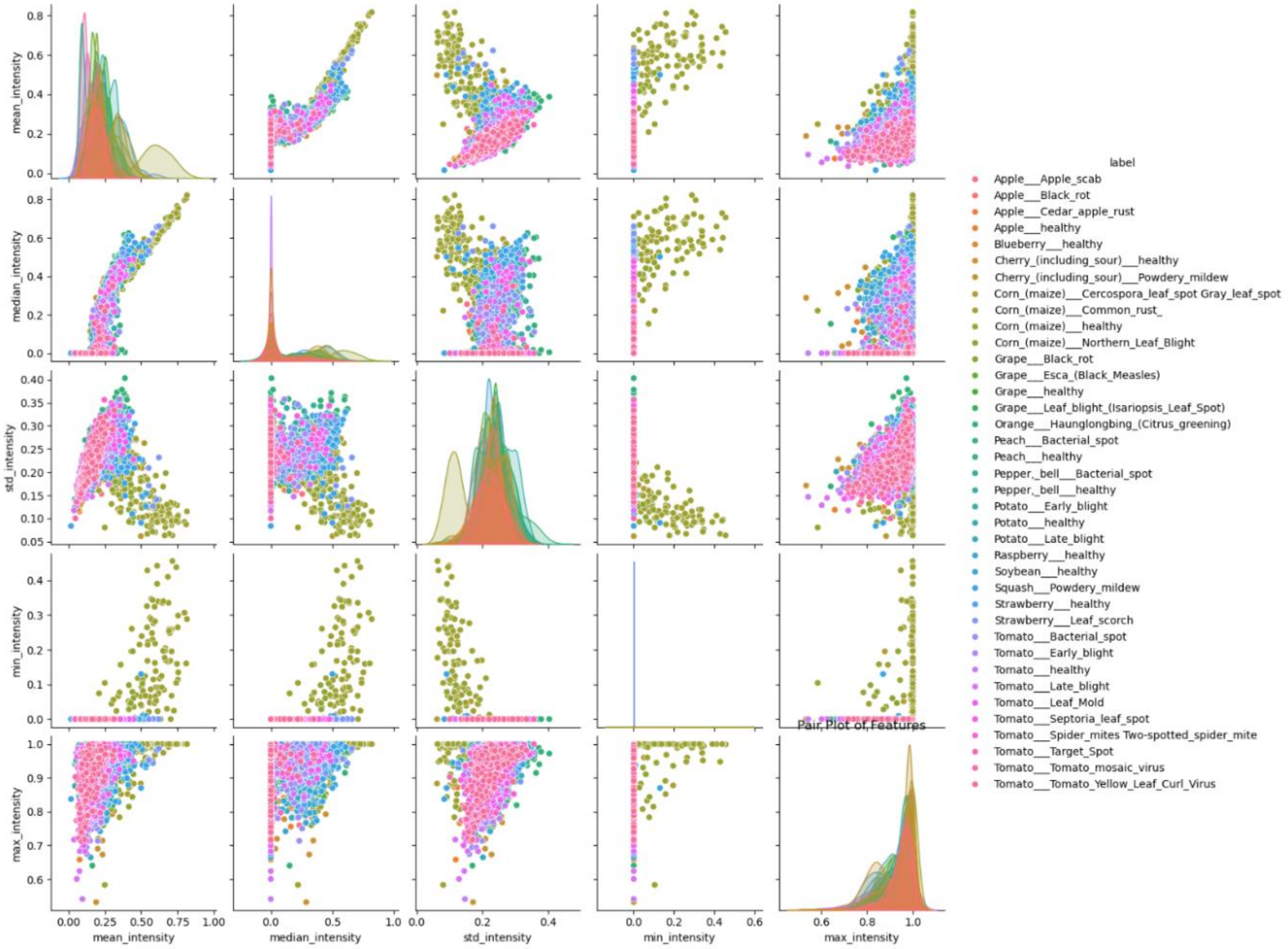
The distinct clustering of points indicates that some species have unique intensity characteristics that can be used for classification. There is significant overlap between some species, suggesting that additional features may be needed to improve classification accuracy.



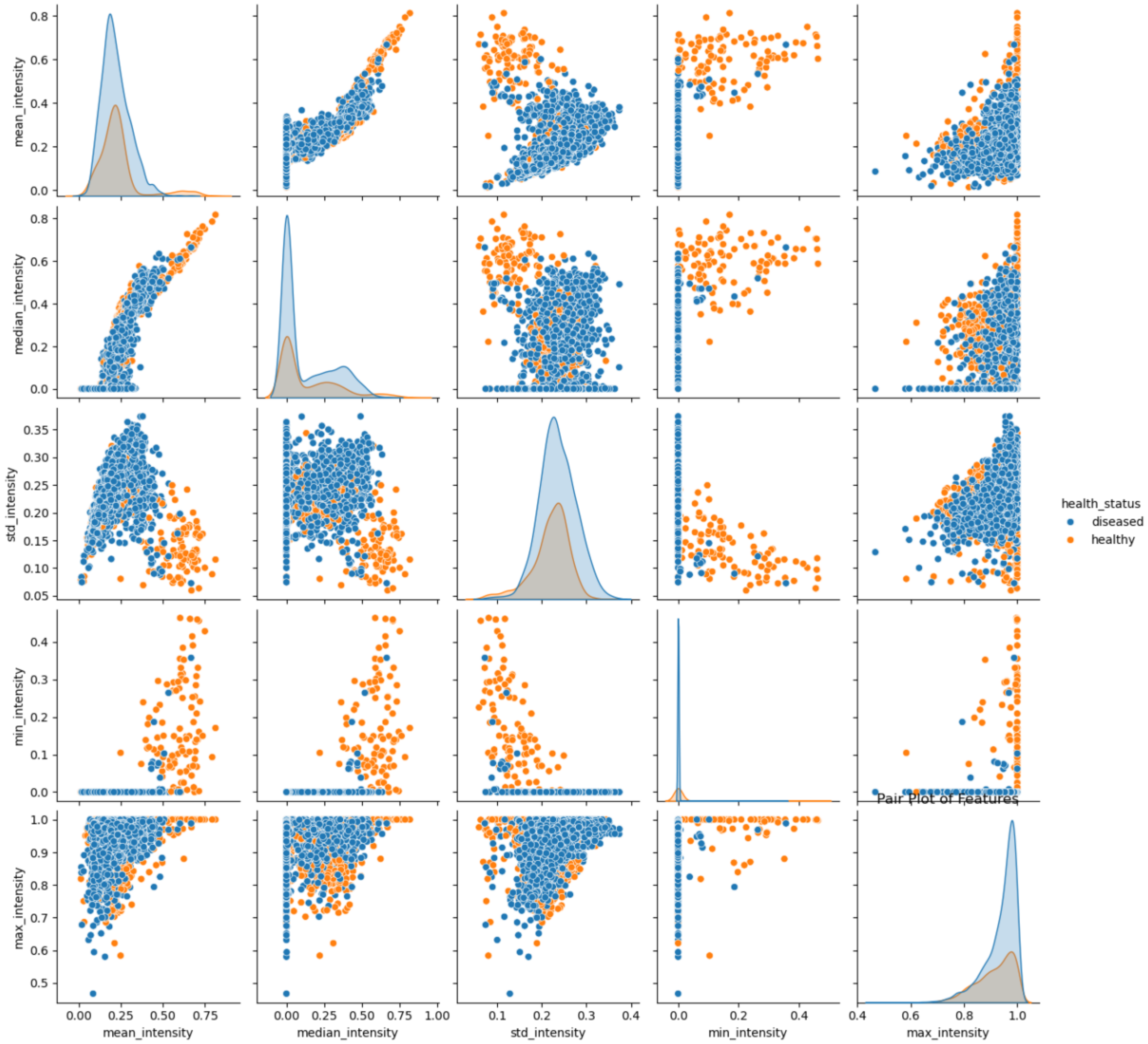
Healthy and diseased plants show distinct intensity patterns, particularly in the mean and median intensities, which can be used to distinguish between the two groups. Despite some overlap, the clear separation in certain features suggests that pixel intensity metrics are valuable for health status classification.



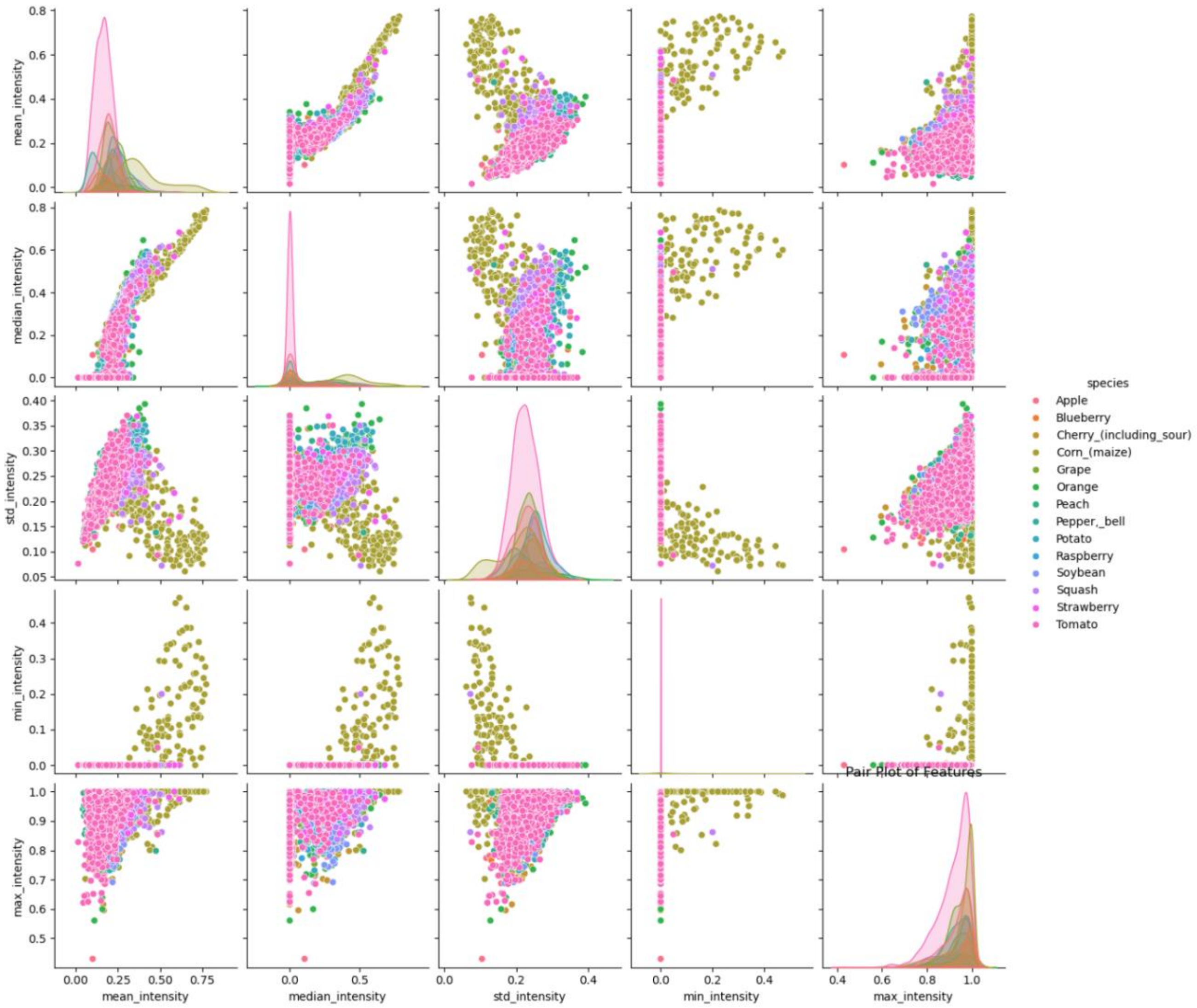
Segmentation enhances the distinction between species by removing background noise, leading to more pronounced clusters. The improved separation of species clusters in segmented images indicates that segmentation is an effective pre-processing step for species classification.



Segmentation helps in better distinguishing between healthy and diseased plants by isolating leaf features, as evidenced by the clearer separation in the plot. The clusters for healthy and diseased plants are more defined compared to non-segmented data, suggesting segmentation significantly aids in disease detection.



The segmentation process aids in highlighting species-specific intensity patterns, making it easier to differentiate between species based on pixel intensity. There is noticeable clustering of data points for each species, indicating that segmented intensity features are effective for multi-species classification.



Analyzing the color histograms for both healthy and diseased leaves can provide insights into the visual differences between them, which can be crucial for model training.

8. Image Quality Analysis

Image quality, in terms of sharpness and brightness, affects the performance of machine learning models. Higher quality images with better sharpness and appropriate brightness levels are more likely to result in accurate models.

Quality Metrics:

- **Blur Score**: Indicates the sharpness of images.
- **Brightness**: Indicates the overall brightness level of images.

Scatter plots of these metrics help in visualizing the distribution of image quality within and across datasets.

9. Model Performance Metrics

Evaluating model performance using metrics such as accuracy and loss over training epochs provides insights into the learning behavior and generalization capability of the model.

Metrics:

- **Accuracy**: Indicates how well the model is performing in terms of correct predictions.
- **Loss**: Indicates the error in the model's predictions.

Plotting these metrics over epochs helps in understanding the model's convergence and identifying any overfitting or underfitting issues.

10. Data Augmentation

To increase the number of images and improve model robustness, data augmentation techniques such as rotation, flipping, stretching, zooming, and color manipulations are used.

Example:

- **Rotation**: Augmenting images by rotating them at various angles.

Data augmentation helps in creating a more diverse training set, which can lead to better model generalization.

11. Conclusion

The analysis of the plant recognition datasets reveals several critical insights and challenges that need to be addressed to improve the efficacy of AI-based plant disease detection systems. The datasets examined include the New Plant Disease Dataset, Plant Disease Dataset, and Plant Village Dataset, with the V2 Plant Seedlings Dataset excluded due to its unsuitability for the project's goals.

There is a notable predominance of tomato images, especially in the New Plant and Plant Disease datasets. This imbalance could skew model training, making the models less effective for other crops. Certain crops, such as squash and raspberry, are significantly underrepresented. This lack of data could lead to models that underperform for these crops, limiting their utility in diverse agricultural settings. While corn and apple are well-represented in the New Plant dataset, they are less prominent in the Plant Village and Plant Disease datasets. This variability suggests that the choice of dataset can significantly impact model performance based on the crop of interest. There is a disproportionate number of diseased images compared to healthy ones across the datasets. This imbalance may cause models to be overly sensitive to detecting diseases, potentially reducing their effectiveness in identifying healthy plants. To mitigate these issues, data augmentation and segmentation techniques can be employed to balance the datasets and improve model robustness. Ensuring high-quality images and leveraging segmented images for better feature extraction will further enhance model performance. This comprehensive analysis provides a foundation for further research and development in plant recognition, contributing to advancements in agriculture and plant pathology.