Project Report: Cassava Leaf Disease Classification Challenge

1. Introduction

Overview of the Challenge and Objectives

The Cassava Leaf Disease Classification Challenge aims to develop a machine learning model capable of accurately identifying diseases affecting cassava plants based on leaf images. This project leverages ensemble learning techniques to enhance classification accuracy and robustness, contributing to sustainable agricultural practices.

2. Data Analysis

Insights from Exploratory Data Analysis (EDA) and Preprocessing

Upon exploration of the dataset sourced from Kaggle, it was found to consist of annotated images categorized into five classes: Cassava Bacterial Blight (CBB), Cassava Brown Streak Disease (CBSD), Cassava Green Mottle (CGM), Cassava Mosaic Disease (CMD), and Healthy leaves. Key insights from EDA include:

- **Class Distribution:** The dataset exhibits varying class distributions, with Cassava Mosaic Disease (CMD) being predominant.
- **Image Quality:** Images vary in resolution and quality, necessitating preprocessing steps such as resizing, normalization, and augmentation to standardize input for model training.

3. Model Development

Detailed Description of Model Architecture and Training Process

Three pretrained CNN models—EfficientNetB0, ResNet50, and MobileNetV2—were utilized to extract features from cassava leaf images. These features were concatenated and fed into a RandomForestClassifier for ensemble learning. Model development involved:

- **Feature Extraction:** Extracting high-level features using pretrained models.
- **Ensemble Learning:** Combining features to enhance classification performance.
- **Training and Validation:** Splitting the dataset into training and validation sets, optimizing model parameters, and evaluating performance metrics.

Challenges Encountered

- **Class Imbalance:** Addressing imbalanced class distributions required careful handling to prevent biased model outcomes.
- **Computational Efficiency:** Balancing model complexity with computational resources influenced the choice of architectures and training strategies.

4. Results and Evaluation

Presentation and Interpretation of Evaluation Results

The ensemble RandomForestClassifier achieved the following results on the validation set:

Classification Report for Validation Data

	Precision	recall	f1-score	support
Cassava Bacterial Blight (CBB)	0.56	0.25	0.34	175
Cassava Brown Streak Disease (CBSD)	0.58	0.32	0.42	347
Cassava Green Mottle (CGM)	0.71	0.09	0.15	371
Cassava Mosaic Disease (CMD)	0.74	0.98	0.84	2133
Healthy	0.55	0.38	0.45	398
Accuracy			0.71	3424
macro avg	0.63	0.40	0.44	3424
weighted avg	0.69	0.71	0.65	3424

- **Accuracy:** The model achieved an accuracy of 71.17% on the validation set, demonstrating competitive performance across disease categories.
- **Precision and Recall:** Varied across classes, with CMD showing high precision and recall, indicating effective disease identification critical for agricultural decision-making.

5. Conclusion and Future Work

Summary of Findings and Limitations

In conclusion, the ensemble learning approach using pretrained CNN models significantly improved the classification of cassava leaf diseases. However, limitations include:

- **Class Imbalance:** Further strategies are needed to mitigate the impact of imbalanced data on model performance.
- **Generalization:** Enhancing model generalization through expanded datasets and additional preprocessing techniques remains a priority.

Suggestions for Future Improvements

Future research could focus on:

- **Temporal Data Integration:** Incorporating temporal data to enhance disease progression monitoring.
- **Advanced Preprocessing:** Exploring advanced augmentation techniques to further diversify the dataset and improve model robustness.