Pathology Detection in Crop Plants

https://github.com/azzadom/FDS-Final-Project

Domenico Azzarito, Guillermo Bajo Laborda, Laura Alejandra Moreno,
Arian Gharehmohammadzadehghashghaei, Michele Pezza

December 29, 2024

Abstract

Early detection of crop diseases is vital for sustainable agriculture. This project classifies apple leaf diseases using machine learning, combining traditional methods like logistic regression and random forests with CNNs. Data preprocessing, including augmentation and feature extraction, further improves performance. The results demonstrate the effectiveness of integrating classical and modern approaches for scalable plant pathology solutions.

1 Introduction

Crop diseases impact agricultural productivity, emphasizing the need for early detection. This project classifies apple leaf diseases using machine learning techniques and the Kaggle dataset "Plant Pathology 2020 - FGVC7." We combine traditional models with CNNs, leveraging preprocessing and feature extraction to build accurate and scalable classifiers.

2 Related Work

Mohanty et al. (2016) demonstrated the effectiveness of convolutional neural networks (CNNs) in detecting plant diseases from leaf images with high accuracy. Their work emphasized the use of large, publicly available datasets, aligning with our approach. Insights from their research guide model design and training strategies, improving performance and scalability in pathology detection for crop plants.

3 Dataset

The dataset used in this study is derived from the Kaggle competition "Plant Pathology 2020 - FGVC7." It comprises 1,821 labeled images of apple leaves, classified into four distinct categories:

- **Healthy**: Leaves exhibiting no visible symptoms of disease.
- **Rust**: Leaves affected by rust-like fungal pathogens.
- **Scab**: Leaves showing scab lesions caused by fungal infections.
- **Multiple Diseases**: Leaves displaying signs of multiple co-occurring diseases.

4 RGB Histogram Analysis

This section details dataset exploration, preprocessing, and feature extraction to classify leaf diseases.

4.1 Exploratory Data Analysis (EDA)

- **Class Distribution:** Leaf categories (Healthy, Rust, Scab, Multiple Diseases) were analyzed for imbalances. The underrepresentation of *Multiple Diseases* prompted data augmentation strategies.
- **Class Characteristics:** Visual inspection revealed uniform colors in healthy leaves, while diseased leaves showed distinct patterns like spots and discoloration, aiding classification.

4.2 Preprocessing and Feature Extraction

- **Image Processing:** Images were filtered, transformed, and cropped using Canny edge detection to isolate leaf regions.
- **RGB Histograms:** Color histograms from the **Red, Green, and Blue (RGB)** channels captured pixel intensity distributions, forming normalized input vectors for classification.

4.3 Insights from RGB Features

- Diseased leaves showed more variability in red and green intensities, while healthy leaves had consistent distributions.

5 Models and Methods

Softmax Regression and Random Forest

We applied two machine learning approaches, Softmax Regression and Random Forest, to classify plant leaf diseases, integrating RGB histograms and Gray-Level Co-occurrence Matrix (GLCM) features.

Feature Extraction

We extracted features from preprocessed images using the following methods:

- **Edge Detection:** Canny edge detection was employed to isolate leaf regions and reduce background noise.
- **Histogram Features:** RGB histograms with 256 bins per channel were computed, concatenated, and normalized to form input vectors.
- **GLCM Features:** Gray-Level Co-occurrence Matrices captured texture properties, such as contrast, homogeneity, and correlation, reflecting spatial relationships between pixel intensities.

Softmax Regression

Softmax Regression served as a baseline classifier. Key steps included:

- **Handling Class Imbalance:** Gaussian noise augmented underrepresented classes, particularly *Multiple Diseases*, to balance the dataset.
- **Standardization:** Features were standardized to ensure consistent scaling across inputs.
- **Model Training and Evaluation:** The model was trained with a multinomial logistic regression solver (1000 iterations) and evaluated using accuracy and classification metrics.

Random Forest

We implemented a Random Forest classifier to enhance classification accuracy and reduce overfitting. The process involved:

- Training a Random Forest model with 200 decision trees.
- Evaluating performance using accuracy metrics and confusion matrices.

Results:
The integration of RGB histograms, GLCM texture features, and ensemble learning methods demonstrated that Random Forest achieved superior performance compared to the baseline Softmax Regression. This highlights the effectiveness of combining feature-based methods with ensemble techniques for plant disease classification.

Convolutional Neural Networks (CNNs)

ResNet-50 Convolutional Neural Network

We employed ResNet-50, a deep Convolutional Neural Network (CNN) characterized by residual connections to address vanishing gradient problems during training. Pre-trained on ImageNet, ResNet-50 was fine-tuned to identify leaf diseases in this study. Its residual blocks facilitated efficient training by allowing gradients to flow through shortcut connections, preserving learning in deeper layers.

Key implementation steps included:

- **Transfer Learning:** The pre-trained weights were leveraged as a feature extractor, with the fully connected layer replaced to output four classes corresponding to the leaf conditions.
- **Freezing Layers:** The initial convolutional layers were frozen to retain general feature extraction capabilities, while the final layers were retrained for the specific classification task.
- **Fine-Tuning:** The final fully connected layer was optimized using the Adam optimizer with a learning rate of 0.001 and weight decay to reduce overfitting.

Class imbalance, particularly for the *Multiple Diseases* category, was mitigated using data augmentation and oversampling strategies. ResNet-50 achieved competitive performance by effectively capturing hierarchical and spatial patterns, even for subtle disease indicators.

VGG-16 Convolutional Neural Network

We also implemented VGG-16, a deep CNN architecture introduced by Simonyan and Zisserman, featuring sequential 3x3 convolutional layers. The pre-trained model, originally trained on ImageNet, was fine-tuned for leaf disease detection.

Key implementation steps included:

- **Transfer Learning:** Pre-trained weights were retained for initial layers, and the classifier was modified to accommodate four output classes.

- **Feature Extraction:** The convolutional layers captured fine-grained patterns such as leaf textures and edges, critical for disease detection.

- **Optimization:** The Adam optimizer with weight decay and a learning rate of 0.001 was used to train the final layers, ensuring stability and convergence.

Despite its effective feature extraction capabilities, VGG-16 faced challenges due to its large size and lack of residual connections, making it susceptible to vanishing gradients. To overcome these limitations, regularization techniques and data augmentation were applied, improving robustness against class imbalances and enhancing performance in detecting subtle disease features.

Evaluation and Comparison

Both CNN architectures—ResNet-50 and VGG-16—achieved strong classification performance, but exhibited differences in computational efficiency and robustness:

- **ResNet-50:** Excelled in handling deeper feature hierarchies due to residual connections, making it less prone to vanishing gradients and better suited for complex patterns.

- **VGG-16:** Provided interpretable and strong feature extraction capabilities but required more computational resources and careful regularization to prevent overfitting.

6 Results and Evaluation

Softmax Regression and Random Forest

Table 1: Performance Comparison: Softmax Regression and Random Forest

Model	Accuracy	Precision	Recall	F1-score
Softmax Regression	77.0%	76.2%	78.3%	77.1%
Random Forest	81.1%	81.3%	82.4%	81.8%

Random Forest outperformed Softmax Regression, achieving an accuracy of 81.1%, compared to 77.0%. Preprocessing steps, including SMOTE and Gaussian noise augmentation, improved Softmax Regression performance. Hyperparameter tuning further optimized both models, demonstrating the importance of data preprocessing and model optimization in enhancing prediction accuracy.

CNN Models: ResNet-50 and VGG-16

The CNN models, ResNet-50 and VGG-16, were fine-tuned on the dataset leveraging transfer learning. Table 2 summarizes the test accuracy achieved by each model.

Table 2: Performance Comparison: ResNet-50 and VGG-16

Model	Epochs	Test Accuracy
ResNet-50	20	83.52%
VGG-16	20	86.81%

VGG-16 outperformed ResNet-50 with a test accuracy of 86.81%, compared to 83.52%. Both models effectively utilized transfer learning, but VGG-16 demonstrated slightly better generalization and robustness during training.

Per-Class Performance

Tables 3 and 4 detail the classification performance metrics for each class.

Table 3: Classification Report: ResNet-50 (20 Epochs)

Class	Precision	Recall	F1-score
Healthy	92.2%	81.0%	86.2%
Multiple Diseases	100.0%	10.0%	18.2%
Rust	79.7%	92.2%	85.5%
Scab	80.4%	90.0%	84.9%

Table 4: Classification Report: VGG-16 (20 Epochs)

Class	Precision	Recall	F1-score
Healthy	94.4%	87.9%	91.1%
Multiple Diseases	50.0%	30.0%	37.5%
Rust	84.1%	90.6%	87.2%
Scab	86.8%	92.0%	89.3%

Both models performed well on the "Healthy," "Rust," and "Scab" classes, achieving high F1-scores. However, the "Multiple Diseases" class remained challenging due to its limited representation, resulting in lower recall values (10.0% for ResNet-50 and 30.0% for VGG-16). These results highlight the importance of addressing class imbalance through techniques like data augmentation and oversampling.

Main Results

The experimental results highlight the following key findings:

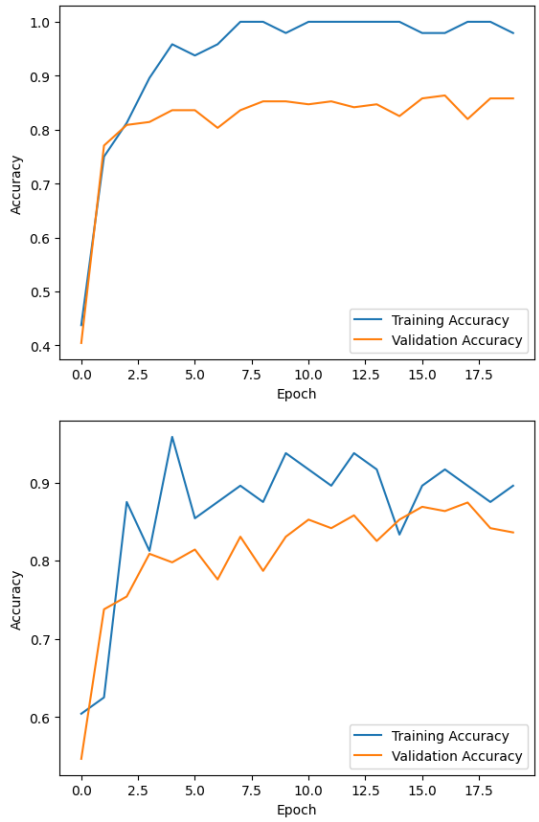

Figure 1: Training and Validation Accuracy Curves for ResNet-50 (up) and VGG-16 (down).

- **VGG-16** achieved the highest test accuracy of **86.81%**, demonstrating robust performance across most classes, particularly "Rust" and "Scab."
- **ResNet-50** followed closely with an accuracy of **83.52%**, leveraging residual connections to handle complex feature hierarchies effectively.
- **Random Forest** delivered strong results, achieving an accuracy of **81.1%**, significantly outperforming Softmax Regression. Its ensemble approach proved effective in capturing patterns from histogram-based features.
- **Softmax Regression** served as a baseline, achieving **77.0%** accuracy, benefiting from data preprocessing techniques like SMOTE and data augmentation.
- Both CNN models struggled to classify the underrepresented "Multiple Diseases" class, with VGG-16 and ResNet-50 achieving recalls of **30.0%** and **10.0%**, respectively. This underscores the need for targeted augmentation and oversampling techniques to address class imbalance.

These results emphasize the advantages of deep learning architectures, particularly VGG-16, for disease detection tasks, while also highlighting the importance of addressing data imbalance to improve performance on rare classes.

7 Conclusions

This study explored a range of machine learning techniques, including traditional classifiers and deep learning models, to detect diseases in apple leaves. Among the tested approaches:

- CNN architectures—**VGG-16** and **ResNet-50**—demonstrated superior accuracy (**86.81%** and **83.52%**, respectively) compared to traditional methods.
- **Random Forest** achieved a competitive performance of **81.1%**, validating its effectiveness as an interpretable and scalable alternative for feature-based classification.
- **Softmax Regression** provided baseline results (**77.0%**), highlighting the need for more sophisticated feature extraction and augmentation strategies.

While CNNs excelled in capturing spatial and texture-based patterns, performance on the "Multiple Diseases" class remained suboptimal due to class imbalance. This challenge underscores the need for further refinement through:

- **Data Augmentation:** Enhancing class diversity using synthetic samples or generative models.
- **Advanced Architectures:** Exploring state-of-the-art networks like EfficientNet and Vision Transformers for improved feature extraction and scalability.
- **Preprocessing Improvements:** Applying precise segmentation and noise reduction to enhance input quality.
- **Optimization Techniques:** Incorporating adaptive learning rates and advanced dropout methods to mitigate overfitting.

Future efforts will focus on refining CNN architectures, potentially designing custom models tailored to this dataset. Additionally, deploying these models in mobile and IoT systems can facilitate real-time disease detection, promoting sustainable and technology-driven agriculture. Such advancements could offer practical tools for farmers, enabling early interventions to protect crops and reduce economic losses.

8 Roles

- Arian Gharehmohammadzadehghashghaei: Exploratory Data Analysis and Performance Evaluation.
- Laura Alejandra Moreno and Guillermo Bajo Laborda: Softmax Regression and Random Forest.
- Domenico Azzarito and Michele Pezza: Convolutional Neural Networks.

References

[1] Shruti Jadon. Ssm-net for plants disease identification in low data regime. *arXiv*, 2020.

[2] Abbas Jafar, Nabila Bibi, and Rizwan Ali Naqvi. Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations. *Frontiers in Plant Science*, 2024.

[3] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016.

[4] Sumaya Mustofa, Md Mehedi Hasan Munna, and Yousuf Rayhan Emon. A comprehensive review on plant leaf disease detection using deep learning. *arXiv*, 2023.

[5] Muhammad Shoaib, Babar Shah, and Shaker EI-Sappagh. An advanced deep learning models-based plant disease detection: A review of recent research. *Frontiers in Plant Science*, 2023.

[6] Jianping Yao, Son N Tran, Saurabh Garg, and Samantha Sawyer. Deep learning for plant identification and disease classification from leaf images: Multi-prediction approaches. *arXiv*, 2023.

[7] Affan Yasin and Rubia Fatima. On the image-based detection of tomato and corn leaves diseases: An in-depth comparative experiments. *arXiv*, 2023.