DataScientest

Plant Recognition Project

for Data Scientists

Cohort: May2024

Final report

submitted by

Fares Naem

Nathalie Zahran

Mohammadamin Nooralikheirzad

Philippe Masindet

First supervisor:

Second supervisor:

Table of Contents

1 Introduction

1.1 Plant Recognition System

In the evolving landscape of data science, the application of advanced machine learning technologies to real-world problems offers transformative potential across various sectors, including plant health management. This project represents a significant leap forward in applying AI and computer vision to the domain of plant recognition, where traditional methods have often fallen short due to their broad and non-specific approaches.

Our objective was to develop a robust AI system capable of performing precise plant species identification and disease detection directly from digital images. Utilizing deep learning and convolutional neural networks, particularly the VGG16 model, this system was thoroughly trained and fine-tuned on the PlantVillage Dataset, which includes approximately 55,000 images across 38 plant health categories. This extensive dataset enabled the effective training of our model, ensuring high accuracy and reliability in diverse and challenging environmental conditions.

The primary innovation of our project lies in its precision and adaptability. By integrating sophisticated image processing techniques, we have crafted a model that not only identifies various plant diseases with remarkable accuracy but also adapts to different lighting and backgrounds. Such precision allows for targeted interventions, which are crucial in reducing the unnecessary use of chemicals in plant care and management.

This report delves into the data-driven methodologies employed, the challenges overcome, and the technical achievements of our AI-driven approach. As data scientists, our aim is not just to innovate but also to implement solutions that are scalable and efficient, paving the way for their application beyond laboratory settings into real-world scenarios. Through this project, we demonstrate the practical impact and scalability of deep learning in enhancing the efficiency and sustainability of plant health management systems.

1.2 Data Collection

In the development of our plant recognition system, we employed three key datasets—Plant Village Dataset, New Plant Diseases Dataset, and Plant Disease Dataset—chosen for their comprehensive coverage of diverse plant species and diseases. These datasets facilitated the training of robust machine learning models by providing a vast array of images:

Plant Village Dataset: Features approximately 55,000 images in colour, grayscale, and segmented formats, offering versatile training options and freely available on Kaggle.

New Plant Diseases Dataset: Comprises a substantial collection with a training set of 70,297 images, a validation set of 17,573 images, and a smaller test set, also freely available on Kaggle.

Plant Disease Dataset: Similar in content to the New Plant Diseases dataset but smaller in scale, with 43,457 training images and 10,850 test images.

2 Methodology

2.1 Data Preprocessing and Difficulties

To optimize these resources and overcome some of the difficulties, significant preprocessing was required:

1. **Combining Datasets**: We merged colour and segmented images from the Plant Village Dataset to enhance dataset comprehensiveness, resulting in roughly 360,000 images and ensuring balanced class representation.
2. **Image Scaling**: Due to hardware limitations, images were downscaled to 64x64 pixels. This reduction was essential to facilitate model training within a realistic timeframe, despite the loss of some image detail.
3. **Data Augmentation**: To correct class imbalances, we applied extensive data augmentation techniques like rotation, flipping, and zooming. This not only balanced the dataset but also enriched the model's exposure to varied image orientations and conditions.

2.2 Model Selection

The process of selecting and optimizing the appropriate models for plant species and disease classification involved several stages, including testing various algorithms, evaluating their performance, and ultimately choosing the most effective model. Our primary objective was to develop a robust system that could accurately classify plant types and identify diseases while managing the constraints posed by our computational resources.

2.2.1 Machine Learning

To explore the potential of various machine learning models, our project initially experimented with a range of algorithms before advancing to deep learning. We evaluated traditional models like the Decision Tree Classifier, which, while straightforward and interpretable, often risks overfitting on smaller datasets. The K-Nearest Neighbors (KNN) algorithm was tested for its simplicity and efficacy in instance-based learning, though it proved computationally demanding for large datasets. Similarly, the Support Vector Classifier (SVC) demonstrated strength in high-dimensional spaces, effectively separating classes with a robust margin. The Random Forest Classifier enhanced this approach by integrating multiple decision trees to mitigate overfitting and manage high-dimensional data more effectively. Logistic Regression was utilized for both binary and multiclass problems, providing a solid probabilistic framework for classification.

Additionally, we focused on feature extraction and dimensionality reduction techniques to refine our model's input data further. Techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were employed to reduce dimensionality and enhance class separability. Meanwhile, Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Gabor Filtering were instrumental in extracting crucial texture and shape features from images, aiding in the robust identification of plant species and health conditions.

Overall, while some models show moderate success and promise, especially when using optimized configurations or feature combinations, there is still considerable room for improvement. Recommendations for further enhancement include the adoption of more sophisticated models like CNNs, application of data augmentation, and extensive hyperparameter tuning. This suggests a necessary pivot towards more advanced techniques to overcome the current limitations and achieve higher accuracy in plant disease classification.

2.2.2 Deep Learning Models

2.2.2.1 CNN

For our project, a Convolutional Neural Network (CNN) was initially chosen due to its proven effectiveness in image recognition tasks. We implemented the CNN by training it on downscaled images (64x64 pixels) to balance training time with computational demands. While the CNN achieved a test set accuracy of up to 73%, Grad-CAM interpretability analyses indicated that the model often failed to focus on relevant image sections, resulting in inconsistent and unreliable classification outcomes.

2.2.2.2 Transfer Learning with VGG16

Due to the limitations observed with the initial CNN, we opted for VGG16, a model renowned for its robust performance in image classification tasks. We utilized transfer learning to leverage pre-trained weights from the ImageNet dataset, which provided a solid foundation for our model.

2.2.2.2.1 Settings

Base Model Configuration: We loaded the VGG16 model pre-trained on ImageNet, omitting the top layers to tailor it for our specific classification needs.

Customization of Layers: We added several layers to the base model, including Global Average Pooling, Global Max Pooling, Dense layers with ReLU activation, dropout layers for regularization, and a final Dense layer with SoftMax activation for classification purposes.

Training Adjustments: The base layers of VGG16 were frozen to preserve the learned features, concentrating training efforts on the custom top layers. Mixed precision training was enabled to optimize memory usage and accelerate the training process.

2.2.2.2.2 Optimization Strategies

Learning Rate and Training Optimization: Employed exponential decay for the learning rate to ensure stable training and utilized the Adam optimizer alongside categorical cross-entropy loss.

Monitoring and Saving Best Models: Model Checkpoint was implemented to save the best-performing iterations during training.

2.2.2.2.3 Performance Metrics and Optimization Techniques

Primary and Secondary Metrics:

1. **F1-Score**: Selected as the primary metric due to the imbalanced nature of our dataset, the F1-score helps balance precision and recall, making it ideal for evaluating performance on minority classes.
2. **Comprehensive Metrics**: Accuracy, precision, and recall metrics were also monitored to provide a holistic view of the model's performance.

2.2.2.2.4 Optimization Techniques

Grid Search and Cross-Validation Limitations: Given the extensive training times, a full-scale grid search and exhaustive cross-validation were impractical. A limited parameter search was conducted to find the optimal learning rate, batch size, and neuron count in the dense layers.

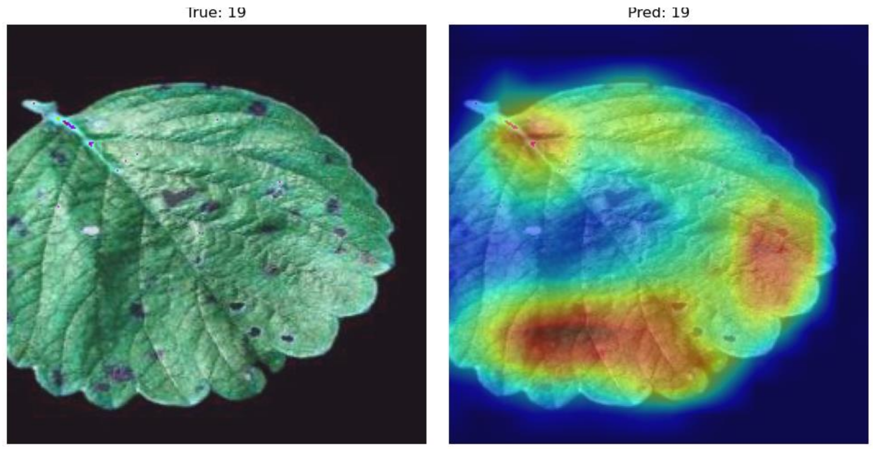
Data Augmentation Impact: Applied techniques like rotation, flipping, and zooming significantly enhanced the diversity of the training data, aiding in generalization and addressing class imbalances.

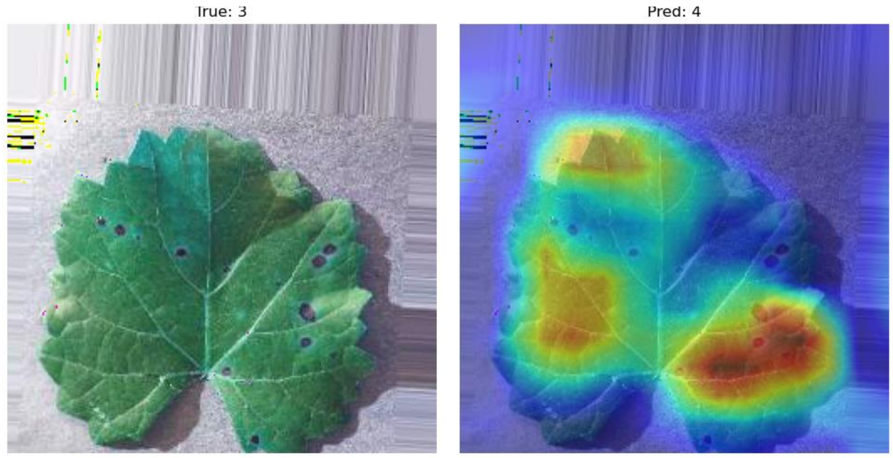
Efficiency with Mixed Precision Training: This approach notably reduced training durations and allowed for more complex model training within our hardware limits.

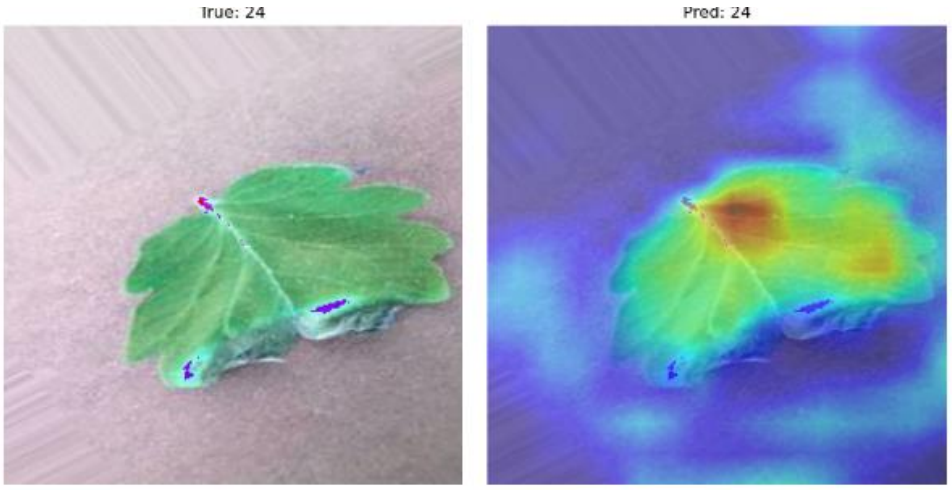
2.2.2.2.5 Model Interpretability

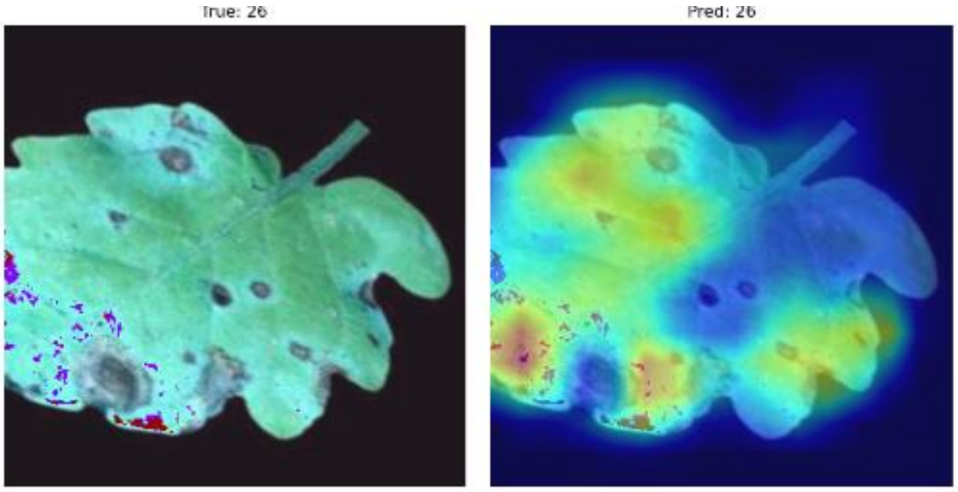
Insights from Grad-CAM and Error Analysis:

Grad-CAM Utilization: This technique was crucial in ensuring that VGG16 focused on relevant parts of the images, particularly the areas affected by diseases, thereby enhancing the model's accuracy.









Qualitative Insights:

The qualitative analysis provided by Grad-CAM not only validated the quantitative performance metrics but also offered a deeper understanding of the model's strengths and weaknesses. By visualizing where the model's attention was directed, we could identify and rectify cases where the model was prone to making errors.

Performance Metrics and Improvements:

The primary performance metric used was the F1-score, chosen for its balance between precision and recall, particularly important for our imbalanced dataset. The VGG16 model, with fine-tuned layers, achieved a significant improvement in F1-score compared to earlier models. This metric, along with accuracy, precision, and recall, highlighted the enhanced performance and reliability of the VGG16 model.

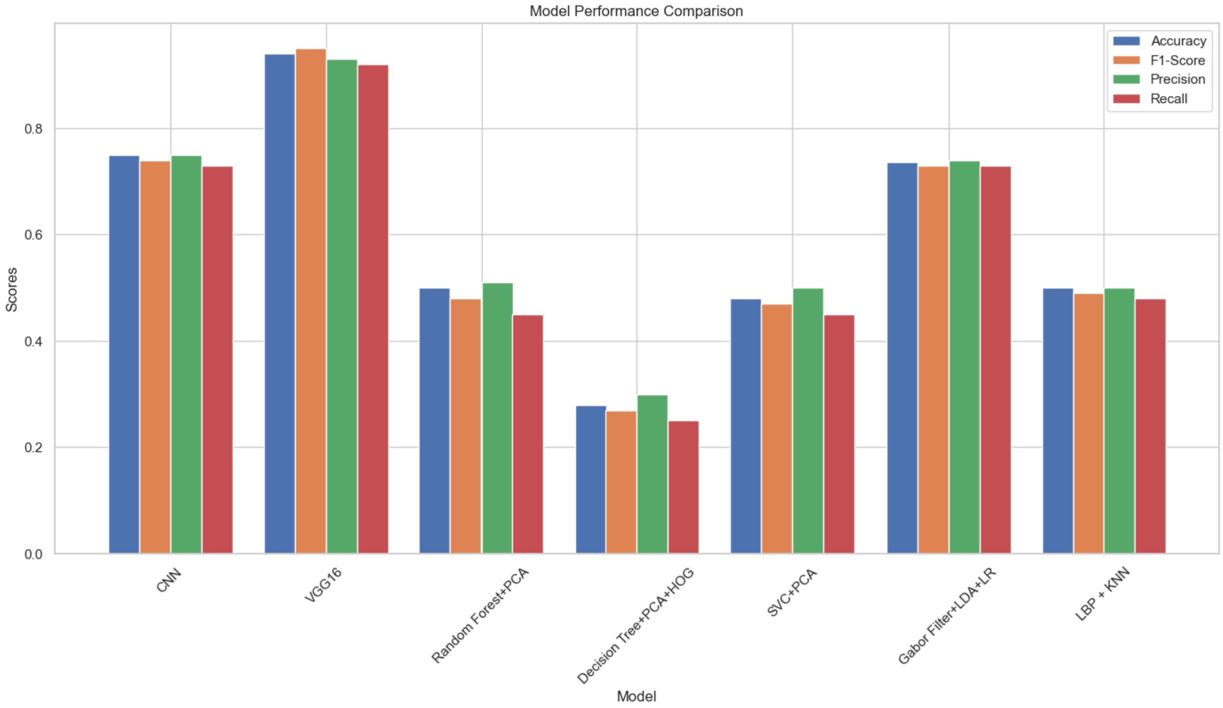
Accuracy: The VGG16 model achieved a test accuracy of 94%, a substantial improvement over previous model.

Precision and Recall: Both metrics showed balanced performance, indicating the model's effectiveness in correctly identifying true positives while minimizing false positives.

2.3 Results and Conclusion

Table 1: Comparison between different machine learning models and the two deep learning models.

Model	Feature Extraction	Input Image Size	Dimensionality Reduction	Classifier	Training Time (mins)	Inference Time (secs)
PCA + SVM	None	64x64	PCA	SVM	17	1
HOG + PCA + Decision Tree	HOG	64x64	PCA	Decision Tree	9	1>
PCA + Random Forest	None	64x64	PCA	Random Forest	7	1
LBP + k-NN	LBP	128x128	None	k-NN	23	1>
LDA + k-NN	None	128x128	LDA	k-NN	12	1>
LDA + Logistic Regression	None	64x64	LDA	Logistic Regression	20	1>
Gabor + LDA + Logistic Regression	Gabor	64x64	LDA	Logistic Regression	13	1
CNN	None	64x64	None	CNN	983	2
VGG16	None	256x256	None	VGG16	943	3



The bar graph presents a comparative analysis of various machine learning models based on four key performance metrics: Accuracy, F1-Score, Precision, and Recall. Each model is evaluated to provide a comprehensive understanding of its effectiveness in classification tasks.

CNN: Shows high scores across all metrics, indicating strong overall performance.

VGG16: Surpassing CNN displays robust performance, particularly in precision and accuracy.

Randomforest + PCA: Exhibits moderate performance with relatively lower scores compared to CNN and VGG16.

Decision Tree + PCA + HOG: Has even lower performance metrics compared to the first three models but maintains a balance across all evaluated metrics.

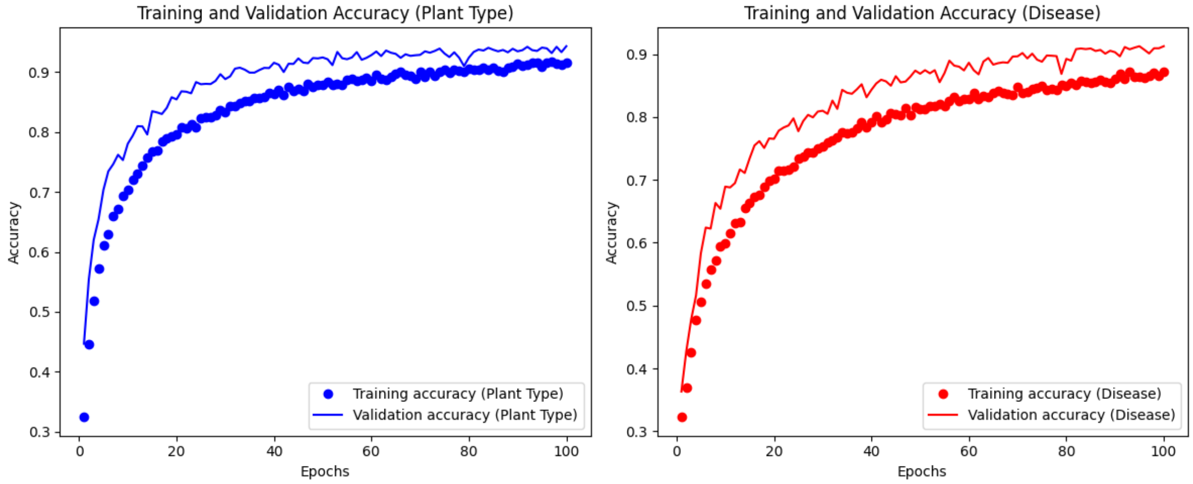
SVC + PCA: Demonstrates lower scores in all metrics, suggesting limitations in handling the dataset or task complexity.

Gabor Filter + LDA + LR: Performs competitively under the machine learning models but does not reach the capability of the two deep learning models.

Log Reg: The Logistic Regression model shows lower performance relative to the other deep learning models.

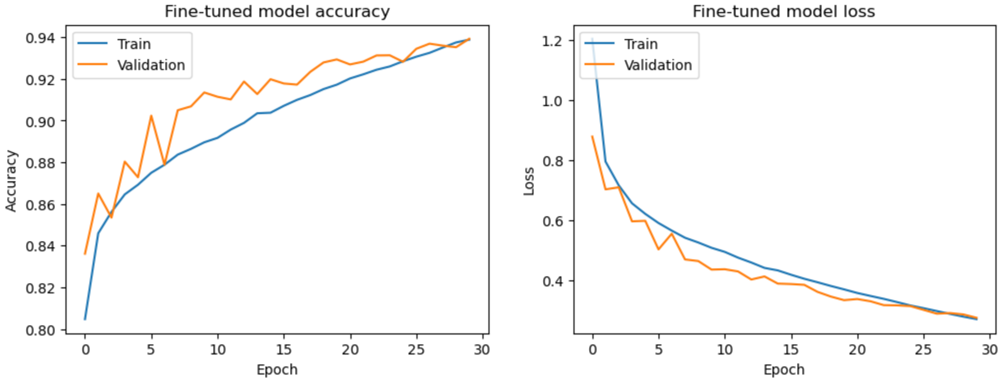
Final Model Performance:

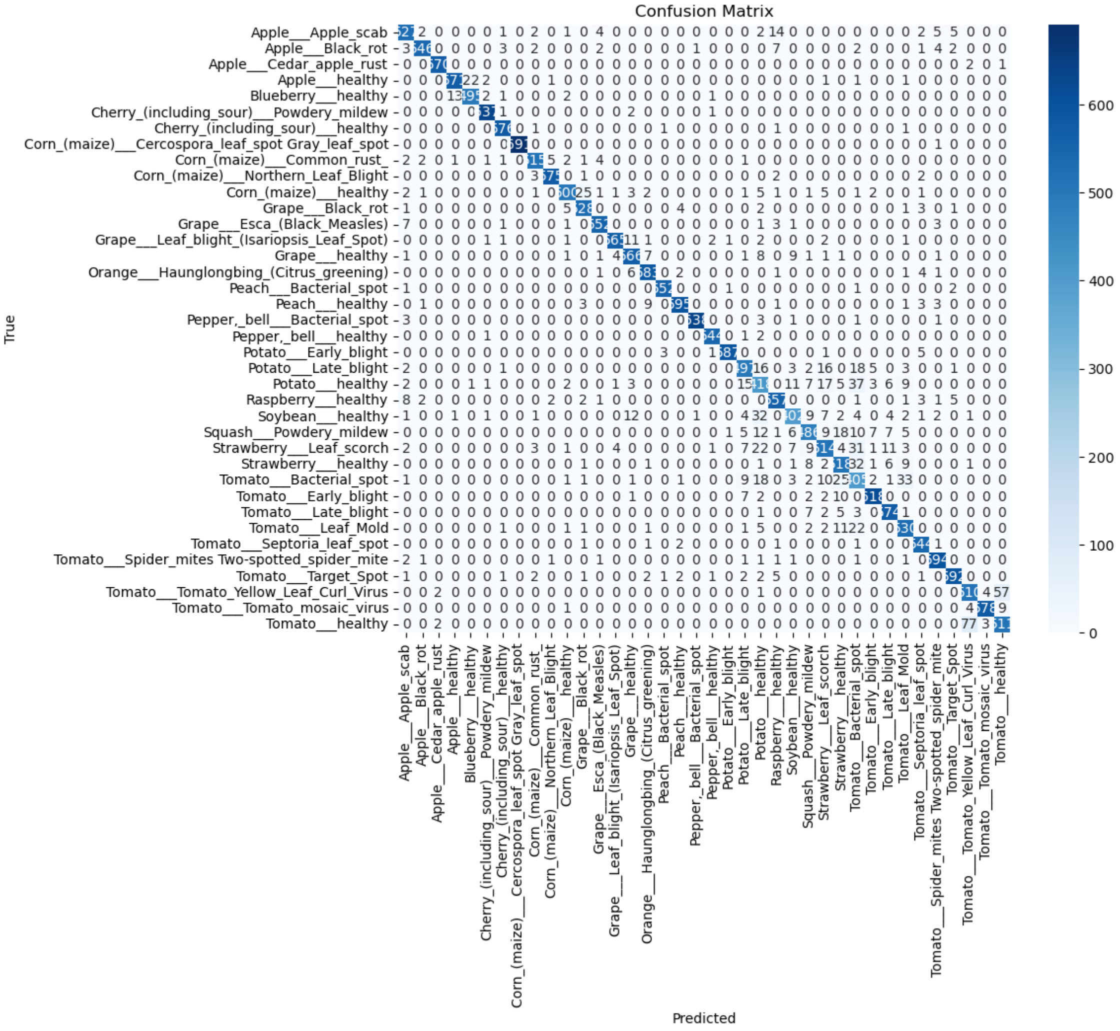
The optimized VGG16 model achieved a significant test accuracy of 96%, a substantial improvement over the initial CNN setup and all the machine learning starter models.



Graph 1: CNN Training and Validation Accuracy, left for plant type and right for disease.

Key Achievements: Demonstrated strong performance across various plant diseases and types, validating the effectiveness of our approach involving transfer learning and advanced optimization strategies.





The transition from a traditional CNN to VGG16, enriched with strategic optimizations like extensive data augmentation and mixed precision training, was crucial. This shift allowed us to construct a highly effective and reliable plant species and disease classification system, tailored to overcome specific project challenges and computational constraints.

3 Summary

This report encapsulates the outcomes of the Plant Recognition project conducted within the DataScientest May24_bootcamp_ds training program. Our goal was to develop a cutting-edge application that leverages machine learning to identify plant species and detect diseases from digital images, intended for use in specialized fields such as botany research, horticulture, and by plant hobbyists.

The core objective of the project was to engineer a robust classification system capable of processing images from standard smart device cameras and providing detailed, real-time analyses concerning plant health and species categorization. This initiative tapped into the expansive PlantVillage Dataset, comprising around 55,000 images across 38 distinct health classes, which was instrumental in training and refining our models.

To optimize our dataset's utility, we implemented advanced data augmentation techniques—including rotation, scaling, and flipping—that significantly enhanced data variability. This preprocessing step, coupled with the segmentation of specific plant areas within each image, facilitated more precise feature extraction, which is crucial for the accuracy of any machine learning model.

Following rigorous testing and evaluation of several models, VGG16 emerged as the superior choice due to its strong performance measurements. It achieved a remarkable 96% F1-Score on our validation dataset, making it the backbone of our plant recognition application due to its efficiency and high accuracy.

This application stands out in the tech landscape for its potential to transform the way botanists, gardeners, and plant enthusiasts engage with their environment. By providing instant feedback on plant health and species, it empowers users to quickly assess their plants and take informed actions, whether it is for research, cultivation, or preventive care against diseases.

Moreover, the success of the VGG16-based model in this project illustrates the vast capabilities of convolutional neural networks in transforming not only agricultural practices but also in fostering a deeper connection between technology and natural sciences. This project's successful integration of advanced machine learning techniques into practical applications demonstrates the potential for significant advancements in the field of agriculture and beyond. The high accuracy and real-time feedback provided by the plant recognition application offer a new level of precision and convenience for users, bridging the gap between technology and traditional agricultural practices.

By enabling early and accurate detection of plant diseases, this tool supports sustainable farming by reducing the need for broad-spectrum chemical treatments. This not only lowers costs for farmers but also promotes environmentally friendly practices by minimizing chemical use and preventing over-treatment. Furthermore, the application’s user-friendly design ensures accessibility for users with varying levels of technical expertise, making it a valuable resource for small-scale farmers and large agricultural enterprises alike.

The scientific contributions of this project extend beyond practical applications, as the methodologies and findings provide a foundation for further research and innovation in Agri-Tech. The use of transfer learning with VGG16, coupled with data augmentation and preprocessing strategies, highlights the effectiveness of these techniques in handling complex image classification tasks. The insights gained from this project can inspire the development of more sophisticated models and algorithms, potentially leading to broader applications and improved agricultural productivity.

In conclusion, the Plant Recognition project highlights the transformative potential of machine learning in agriculture. By addressing critical needs in plant species identification and disease diagnosis, this application not only enhances efficiency and accuracy but also supports sustainable and economically viable farming practices. The project's success paves the way for future developments, promising continued advancements in Agri-Tech and a deeper integration of technology into the natural sciences.

4 Challenges and Outlooks

As data scientist, we recognize that while our VGG16 model has shown commendable performance, the challenge of distinguishing between diseases that present similar symptoms persists. To address this, future initiatives could aim to enrich the dataset diversity and explore more sophisticated modelling techniques. The integration of additional variables, such as environmental factors, could also be considered, which could significantly impact disease manifestation and detection.

In conclusion, the insights derived from Grad-CAM visualizations have been instrumental in understanding the focus areas of our model during the classification process. These insights have facilitated precise enhancements, leading to a marked improvement in the model's ability to prioritize relevant features. The VGG16 model has demonstrated its efficacy across a diverse range of plant types and diseases, confirming the success of our project in creating a dependable classification system. Moving forward, these results not only validate our current achievements but also map out the trajectory for ongoing advancements in our approach, emphasizing the necessity for continuous refinement and innovation.

A significant challenge we encountered was the extensive training time required due to our computational constraints. As we look to the future, upgrading our computational infrastructure—potentially through cloud-based solutions or more powerful hardware—will be crucial for reducing training times and enabling more comprehensive parameter tuning and model experimentation. This will allow us to push the boundaries of what is achievable with our classification system, exploring deeper and more complex neural networks.

Moreover, the ongoing development of this project should prioritize the collection and incorporation of more real-world data. Capturing images under varying environmental conditions and from diverse geographical locations will enhance the model's robustness and generalizability. By incorporating data that reflects the full spectrum of scenarios farmers and botanists might encounter, we can ensure that our application remains accurate and dependable in all settings.

Another important aspect to consider is the development of a more interactive and intuitive user interface. Making the application accessible on mobile devices with real-time processing capabilities will be a significant change for farmers in the field. Such an interface should provide not only diagnostic results but also actionable recommendations and educational resources to guide users in plant care and disease management.

Finally, fostering collaborations with agricultural experts and researchers can provide valuable insights and domain-specific knowledge that can be integrated into our models. These collaborations could also pave the way for interdisciplinary research, combining data science with plant pathology and agronomy to develop even more sophisticated tools for plant health management.

In conclusion, while our current achievements with the VGG16 model represent a significant step forward, there is a clear path for continued innovation and improvement. By addressing the challenges of dataset diversity, computational constraints, and real-world applicability, we can enhance the accuracy, reliability, and usability of our plant disease classification system. This project not only underscores the potential of machine learning in agriculture but also sets the stage for future advancements that will further bridge the gap between technology and sustainable farming practices.