Crop Disease Detection and Risk Prediction using CNN and Weather Data

A Machine Learning Project Report

Submitted by:

Pranita Mandrekar

IMSc Data Science, 1st Year

Department of Data Science

Goa Business School, Goa University

Under the Guidance of:

Prof. [Sanket Mhamal]

Goa Business School, Goa University

Date of Submission: November 2025

Academic Year: 2025–26

Goa Business School

Goa University

Taleigao Plateau, Goa – 403206

Abstract

Agriculture plays a crucial role in the Indian economy, and plant diseases cause significant loss in yield and quality. Farmers often struggle to identify crop diseases at an early stage, leading to delayed actions and reduced productivity. This project focuses on the development of a hybrid Machine Learning system that combines image-based disease detection using Convolutional Neural Networks (CNN) with weather-based disease risk prediction using Random Forests. The system also integrates a user-friendly Streamlit interface for real-time predictions. By combining visual and environmental data, the project aims to provide farmers and researchers with an efficient, scalable, and intelligent crop disease diagnosis tool.

1. Introduction

The agricultural sector forms the backbone of India's economy. However, it faces numerous challenges, including pest attacks, nutrient deficiencies, and plant diseases. Timely identification and management of crop diseases can significantly improve yield and sustainability.

Traditional disease diagnosis relies heavily on manual inspection, which is time-consuming and often inaccurate. With advancements in Artificial Intelligence (AI) and Deep Learning (DL), particularly in Computer Vision, automated plant disease detection has become a reality.

In this project, we integrate two dimensions of prediction:

1. Image-based disease classification using a Convolutional Neural Network (CNN).
2. Weather-based risk assessment using a Random Forest model.

The final model, deployed via a Streamlit dashboard, enables the user to upload a leaf image and select a location to predict both disease type and infection risk level.

2. Literature Review

Several studies have been conducted on plant disease detection using deep learning models. Early works utilized traditional image processing techniques such as feature extraction using texture and color histograms. However, with the introduction of deep neural networks, models such as AlexNet, ResNet, and MobileNet have achieved high accuracy in plant disease classification.

In 2018, the PlantVillage dataset was introduced, which became the benchmark for agricultural AI research. The dataset consists of over 50,000 images of healthy and diseased leaves across various plant species.

Previous models, however, primarily focused on visual data. This project adds novelty by integrating meteorological data (temperature, humidity, rainfall, pressure, and wind speed) into the prediction process. By combining CNN-based image analysis with Random Forest-based environmental modeling, the system provides a holistic understanding of crop disease risk.

3. Methodology (Part 1) — Dataset and Preprocessing

3.1 Dataset Source

The dataset used for this project was obtained from Kaggle's repository "*Plant Disease Dataset*" by Emma Rex. It contains thousands of labeled images of crop leaves categorized into multiple disease classes, including tomato, potato, corn, and grape diseases.

3.2 Data Preprocessing

The raw images were downloaded directly using the Kaggle API in Google Colab. Preprocessing steps included:

- Resizing all images to 224×224 pixels.
- Splitting the dataset into 80% training and 20% testing.
- Organizing images into respective subfolders for each class.

This process was automated through Python scripts and verified by visual inspection of random samples. The clean data folders were stored in the '/data/images/' directory of the GitHub repository.

3.3 Tools Used

The main tools used for this stage were:

- Google Colab – for running preprocessing scripts and model training.
- GitHub – for version control, collaboration, and code management.
- Kaggle API – for automated dataset downloading.

4. Methodology (Part 2) — Model Development

4.1 CNN Model for Image Classification

The CNN model was built using TensorFlow and Keras libraries. Transfer learning was applied using the MobileNetV2 architecture, pre-trained on ImageNet. The network was fine-tuned for crop disease classification with softmax output for multiple disease categories.

The model was trained for 10 epochs, and the final model file was saved as `cnn_model.h5`.

4.2 Weather Risk Prediction Model

A synthetic weather dataset was generated using Python's random module to simulate environmental parameters such as temperature, humidity, rainfall, and pressure. The data was labeled as "High" or "Low" disease risk based on predefined conditions.

A Random Forest Classifier was trained on this dataset to predict risk levels. The model was saved as `weather_model.pkl`.

5. Integration and Implementation

Both models were integrated into a single workflow using Python. The CNN model predicts the disease from the image, and the Random Forest model predicts the risk based on live weather data fetched from the OpenWeatherMap API.

A Streamlit dashboard was created to make the system interactive. Users can:

- Upload a crop leaf image.
- Enter a city name.
- View disease name, confidence score, and risk level.

This user interface enhances accessibility and practical use for farmers and agricultural analysts.

6. Results and Discussion

The CNN model achieved high accuracy in distinguishing healthy and diseased leaves. The weather model successfully predicted risk levels consistent with environmental patterns.

Sample output:

```
Image Prediction: Tomato Bacterial Spot (Confidence: 0.91)
Weather Data: {Temp: 30.4°C, Humidity: 85%, Pressure: 1002 hPa}
Predicted Risk: High
```

The integration of weather conditions improved contextual prediction, helping users understand both the current disease and the potential spread risk.

7. Personal Contribution (Pranita Mandrekar)

As part of this group project, I contributed primarily to two major components — **data preprocessing** and **model integration**.

7.1 1. Image Preprocessing

- Downloaded and prepared the image dataset using Kaggle API.
- Resized and split the dataset into training and testing sets using Python scripts.
- Structured and verified folders in Google Colab for clean model input.

7.2 2. Integration and Streamlit

- Worked on building the Streamlit dashboard interface for integrated predictions.
- Connected the CNN image classifier and weather-based Random Forest model for joint inference.
- Tested the system functionality to ensure smooth predictions.

7.3 3. Collaboration Tools

- Used GitHub for version control, creating branches, pushing updates, and managing pull requests.
- Utilized Google Colab for running all preprocessing and integration code collaboratively with teammates.

Through this, I learned to manage end-to-end ML workflows, integrate multi-model systems, and use professional tools for team-based projects.

8. Conclusion and Future Scope

This project successfully demonstrates an end-to-end approach to crop disease diagnosis and risk prediction by integrating computer vision and environmental modeling. The results show that combining image and weather data improves overall prediction reliability.

Future Scope

- Expanding the dataset with real-time field images and local sensor data.
- Deploying the model as a mobile application for farmers.
- Adding real-time alert systems for high-risk disease conditions.

References

1. Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*.

2. Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*.

3. Kaggle Dataset: `https://www.kaggle.com/datasets/emmarex/plantdisease`

4. OpenWeatherMap API: `https://openweathermap.org/api`