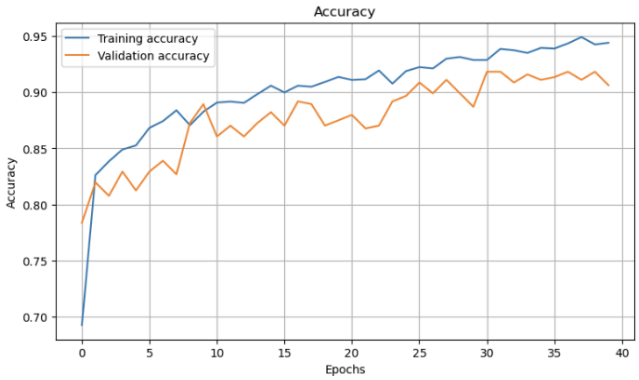
Corn Disease Detection and Identification

Group 27
Patrick LaVallee, Data Science Undergraduate
5/4/23

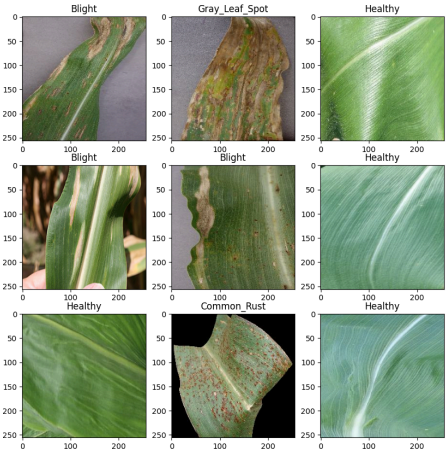
ABSTRACT
A dataset consisting of pictures of corn diseases(Blight, Grey Leaf Spot, Common Rust) was found on kaggle along with some previous works. The data has been preprocessed by scaling the photos down to between values of 0 and 1 to make training data more efficient. The data has also been partitioned into training, validation, and testing data. The chosen models that will be trained are going to be a CNN model using multiple Conv2D layers with varying amounts of filters, followed by max pooling layers along with a MobileNetV2 model. The models will then be trained. Following the training, the models will be used with the testing data to find accuracy and the data will be displayed.

Introduction

Corn is one of the most important cereal crops globally and its productivity is significantly affected by various diseases. Early detection and management of these diseases are crucial for ensuring optimal yield and quality. The goal of this project is to create an image classification model that can take an image of a corn leaf and return back if it is healthy, or what disease it has. There have been other works on this such as the work on kaggle by Imran MohanJka, which gives a 92% classification accuracy. It uses a CNN image classifier using multiple Conv2D layers and max pooling layers. The Conv2D layers have varying filters, which could be played around with to see if there is a better model to train than this original one. This is a pretty basic and simple CNN model.



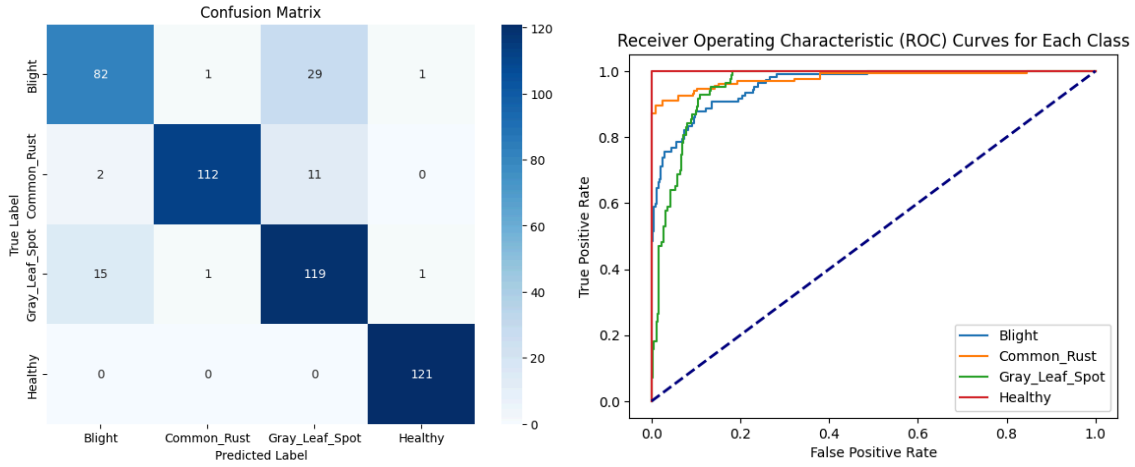
Dataset: The dataset for this project consists of images of corn plant leaves sorted into four different diseases: Common Rust, Gray Leaf Spot, Blight, and Healthy. The dataset has been curated from the PlantVillage and PlantDoc datasets, with certain images removed for relevance. It comprises 1306 images of Common Rust, 574 images of Gray Leaf Spot, 1146 images of Blight, and 1162 images of Healthy leaves. The images are sized to be 256x256x3 (For RGB), and then data is put into batches of 32 and being partitioned 70%/20%/10% (Training, Validation, Testing). The data is also scaled down to values between 0 and 1, improving training efficiency. An example of a sample is a scaled down image with a label of 0, 1, 2, or 3, showing which disease the plant has.



Tasks performed: After preprocessing the data the main objective is to define the models that will be used. Much like the base-line approach, a simple CNN model with alternating conv2d layers, increasing filters each time, and max pooling layers followed by two dense layers will be used. Also, an untrained MobileNetV2 model will be made, adding some custom dense layers for our class amount. MobileNetV2 is a 53 layer deep CNN that typically identifies objects, but if with it untrained, it can be trained to however many classes are needed. With the MobileNetV2 the images must be resized to 224x224 to fit the model's expected input shape. With these two models made, they're then trained. The basic CNN model is just be trained for 30 epochs, while the MobileNetV2 is trained for 100 epochs, but includes early stopping, this is to account for how much deeper the model is.

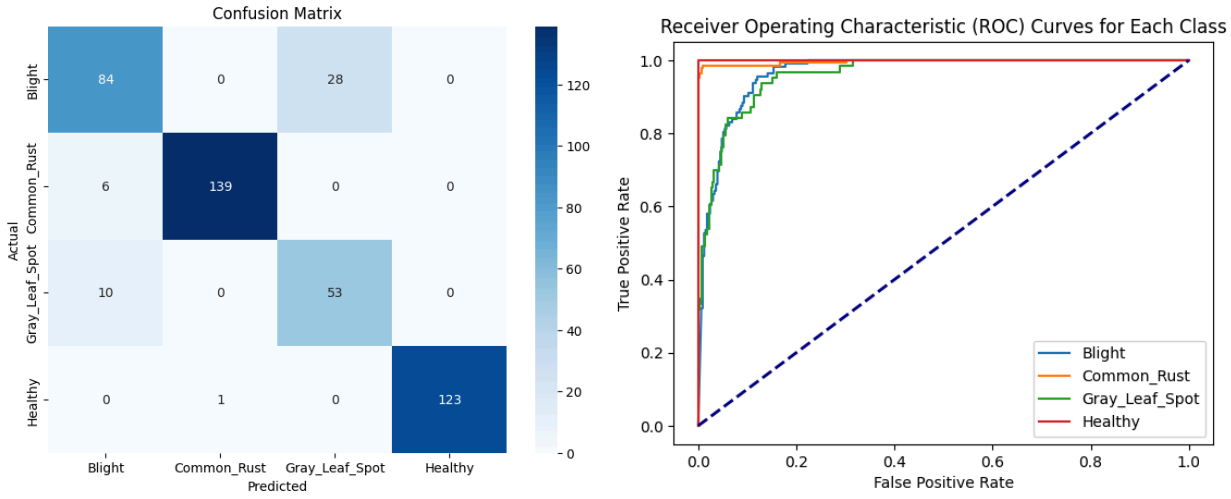
Results and Discussions:

Simpler CNN Model



The simpler CNN model obtained an accuracy of 88%, which is not too bad compared to the baseline of 92%. As seen on the confusion matrix above, the biggest problem for the model was distinguishing between Gray_Leaf_Spot and Blight. This could be due to the smaller sample size of Gray_Leaf_Spot, having not even half the amount of data as the other 3 labels. Blight and Gray Leaf Spot share a region of the genome associated with resistance and can sometimes be confusing to distinguish between just based on looks (Cooper). Although it's not perfect, being able to distinguish between most Gray_Leaf_Spot and Blight is still a pretty good result.

MobileNetV2 model



This MobileNetV2 model trained to obtain 90% accuracy, which still isn't up to par with the base-line related works model, but still pretty good. Again the same issue is happening between Blight and Gray_Leaf_Spot, although it's still able to distinguish most of them correctly.

Looking at both ROC curves, the healthy has a perfect 1.0 for area under the curve on both graphs. Having this good of identification for healthy corn makes the idea of implementing a binary model very smart. Implementing a binary model then focusing down the data even more into just 3 diseased corn classes could allow the model to focus more on distinguishing between the diseased classes, maybe giving a higher accuracy.

References

Cooper, J.S., Balint-Kurti, P.J., & Jamann, T.M. (2018). Identification of quantitative trait loci for Goss's wilt of maize. *Crop Science, 58*(3), 1192-1200. DOI: 10.2135/cropsci2017.10.0618

Mohankjha. (2023). Corn & maize disease detection. *Kaggle.* https://www.kaggle.com/code/mohankjha/corn-maize-disease-detection

Contributions:
My partner did 50% of the proposal then dropped the class

Patrick LaVallee 100% of the rest