2025 Term Paper Project HW-1

Kuan-Chen, Chen

For this project, I choose **Tomato leaf disease detection on Kaggle, this goal of this project is to make me learn more about machine learning, data science techniques on image data, and comparing the different methods on same dataset and compare which kinds of method can make our predictions have higher accuracy.**

So first of all, we take a look at our dataset, we get 10 different category, and each category represents a kind of disease of tomato leaf, we got Tomato mosaic virus, Target Spot, Bacterial spot, Tomato Yellow Leaf Curl Virus, Late blight, Leaf Mold, Early blight, Spider mites, Two spotted spider mite, Tomato healthy and Septoria leaf spot. Each category have 1000 training image and 100 test image.

Regardless to the low numbers of data, so at first we have to make our dataset bigger, by data augmentation, we can flip the data by horizon, vertical, no only but also rotate zoom in zoom out, and this can lead us to having more data, may can let we having higher accuracy and also prevent from overfitting. After all that we split the test dataset into test and validation set after all this we can start to train our data.

So this project is divided into three parts, classification, regression, clustering, and the classification part is that Ill use CNN to classify the image and compare it with different CNN techniques.

Later on we got regression, because we don't have numerical data, so we have to create by our self , so I decide to create a model to detect how much brown particles is in the image, if the image contain more green particles, it means that the leaf is more healthy, and I'll add a scale representing the what disease the leaf gets after adding the percentage and the scale, theres a grade represent "How" health the leaf is, and after we get all the score, we can do the regression part by implement the test data, and the model will give a score output.

Last we got cluster to do, ill use the output of part 2, divided the score into three different part, and there will be three health level for the data, after that ill cluster the test data.