PLANT DISEASE DETECTION FOR MAIZE

Mapamela Wendy
University of Pretoria
u23970911
u23970911@tuks.co.za

Radiokana Boikanyo
University of Pretoria
u16097492
u16097492@tuks.co.za

ABSTRACT

Farmers in Ghana and West Africa face significant challenges controlling and detecting plant diseases in maize which leads to low harvests. To avoid further losses, this project focuses on developing lightweight models for early detection and classification of maize plant diseases. The dataset used is obtained from University of Cape Coast, Ghana. An in-depth analysis of the dataset was conducted, including data quality assessment and pre-processing techniques necessary for effective model development. The analysis revealed an imbalance in class distribution which could potentially bias the model performance. While noise removal techniques were explored, they were ineffective due to significant information loss. Data augmentation was observed in certain classes, while duplicates were identified and thus excluded. Additionally, image sizes varied across classes, with some exhibiting lower resolutions. The paper outlines the critical steps taken to understand and pre-process the dataset, laying foundation for model development.

Keywords

Maize, plant disease, noise removal, data augmentation, images

1. INTRODUCTION

Maize is a crucial crop for farmers in Ghana and West Africa, but recent low yields due to plant diseases have become a significant concern. The project aims to develop lightweight models for early detection and classification of plant diseases in Maize. The dataset used in this project was obtained from University of Cape Coast, Ghana. Sourced from various platforms including but not limited to, Lacuna, Plant Ninja, Zenodo, Roboflow amongst others. The data was received in compressed zip folders, totalling 3.98 gigabytes in size encompassing mainly image data in JPEG format. Additionally, it is worth noting that some folders within the dataset contain non-image files, such as JSON and XML, which contain metadata about the images. To successfully develop lightweight models, it is crucial to understand the characteristics of the dataset, its structure and to uncover insights through exploratory data analysis (EDA). This paper details the data analysis techniques employed on the dataset to identify data quality issues, understand data distribution and patterns, thus enabling effective data pre-processing and ensuring that the data used for analysis and modelling are accurate and reliable. The paper is organised as follows; section 2 discusses the background, section 3 discusses data overview, section 4 details data pre-processing techniques and section 5 concludes this paper.

2. BACKGROUND

Plant diseases are caused by various factors, including pathogens and climate change amongst others. Diseases affecting maize can vary significantly in their characteristics. Therefore, it is crucial to understand the distinctions and characteristics of each disease. Table 1 provides a visual representation of the diseases to be analysed in this project, which include; Common Rust, Northern Leaf Blight, Lethal Necrosis, Streak Virus and Cercospora Leaf Spot. It can be observed that there is variation in background, orientation and brightness. Majority of the classes are captured as single leaf while streak virus has multiple leaves and, in some cases, other plants in the background.

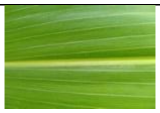
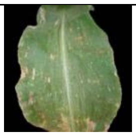
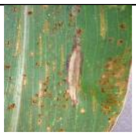
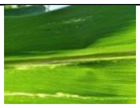
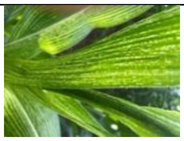
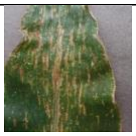
Healthy	
Common Rust	
Northern Leaf Blight	
Lethal Necrosis	
Streak Virus	
Cercospora Leaf Spot	

Table 1:Images of healthy and disease classes

3. DATA OVERVIEW

The dataset for plant disease detection in maize primarily consists of images; categorised into five disease classes: Common Rust, Lethal Necrosis, Cercospora Leaf Spot, Northern Leaf Blight, Streak Virus, along with a class representing healthy maize images. The dataset is labelled using different folders to separate the classes. The distribution of the images across the classes is illustrated in figure 1, which indicates that the healthy class has the highest quantity of images, whilst Northern Leaf Blight disease has the least number of images. Table 2 below details the data distribution percentages across the classes. It can be observed that there is an imbalance within the datasets. Imbalance in data can pose challenges such as biasing the model towards the majority class, leading to poor performance on the minority class. It can also result in misleadingly high accuracy if the model simply predicts the majority class for all instances [1]. In order to address these issues, during model development, various techniques to balance the dataset will be explored.

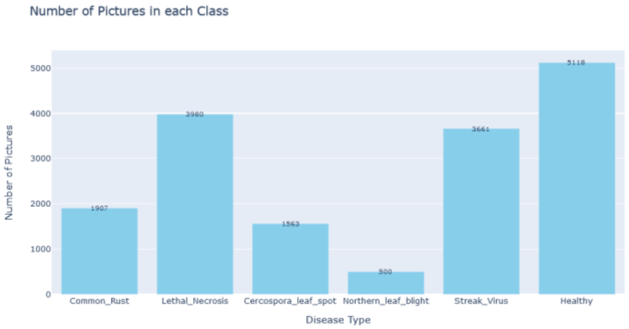


Figure 1:Image distribution

Cercospora Leaf Spot	Common Rust	Lethal Necrosis	Northern Leaf Blight	Streak Virus	Healthy
11%	24%	9%	3%	22%	31%

Table 2: Data distribution percentages

4. DATA PREPROCESSING

4.1. Noise Removal

In an effort to enhance the quality of the dataset by reducing noise in the images, the effectiveness of background removal was examined. It was observed that certain images such as those in healthy class (refer to table 1), were captured as close-ups, while others had background noise, such as multiple plants visible in the background, as seen in the Streak Virus class (refer to table 1). Common Rust images already had their backgrounds removed. Figure 2 illustrates the before and after effects of background removal, revealing a notable loss in image content, resulting in information loss. Therefore, this approach will not be adopted.

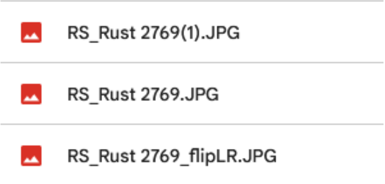
4.2. Data Quality

A. Data Augmentation

After closely observing the various classes, it was noted that certain images had already been augmented to increase the volume of data within the respective classes. These augmentation techniques involve but not limited to, 30, 90, 180 and 270-degree flips which was inferred from the naming convention of the images.

B. Duplicates

The Common Rust class contains 22 (1.15%) duplicates, as indicated by the naming convention below, where images have an extension (1).JPG. Duplicates in Cercospora Leaf Spot class were also observed, this class only contained 4 duplicate images. The identified duplicates are immaterial and will therefore be excluded from model training process.


RS_Rust 2769(1).JPG

RS_Rust 2769.JPG

RS_Rust 2769_flipLR.JPG

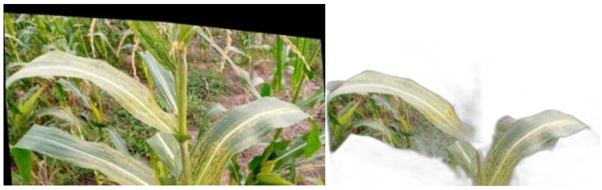


Figure 2: Before and after noise removal

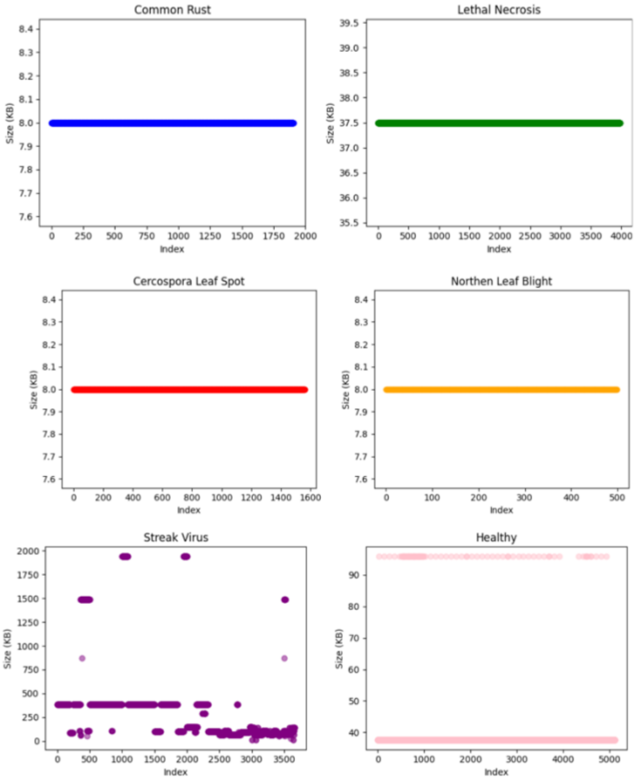


Figure 3: Image sizes for different classes

C. Other Data Formats

As previously mentioned, certain classes have XML and JSON files containing metadata for the images. This metadata typically includes details such as image name, height, width and other relevant information. However, since majority (4/6) of the classes do not have metadata, these files will not be taken into account during model development.

4.3. Image Quality Sizes

The pixel data of the images for each class was extracted and converted to a kilobyte scale using the following method:

For example, if pixel is (256x256):

- 256 x256=65536 bits
- 65536/1021=64 kilobits
- 64/8=8kilobytes

Figure 3 illustrates the various image sizes in kilobytes that were found in the respective classes. It is notable that Cercospora Leaf Spot, Common Rust and Northern Leaf Blight had the lowest image sizes of 8KB, indicating the lowest image resolution. This is also evident in the visual representation of these three classes in table1, where the picture quality appears to be lower compared to healthy, Lethal Necrosis and Streak Virus classes. The Streak Virus and healthy classes have multiple image sizes. To ensure uniform dimensions for all images, during model development image resizing will be implemented.

The link to the code can be found here: https://github.com/vani-radiokana/MIT808-eda-plant-disease-detection-in-maize

5. CONCLUSION

In order to uncover insights and understand characteristics of the dataset. This paper provided a comprehensive overview of the dataset and the essential steps taken to pre-process the data. Through the analysis, various insights were uncovered, including imbalance in the dataset and data quality issues such as duplicates. Despite exploring noise removal techniques, they proved to be ineffective due to information loss. Moreover, variations in images sizes across classes were observed, necessitating the need for image resizing. Data augmentation such as flipping and rotation of images was observed in certain classes.

REFERENCES

[1] M. N. M. Salleh, R. Saedudin, K. Hussain, M. F. Mushtaq and H. Ali, “Imbalance class problems in data mining: a review,” *Indonesian Journal of Electrical Engineering and Computer Science* , vol. 14, no. 3, pp. 1560-1571, 2019.