Yellow Rust Data Set, used to differentiate the severity of the disease
https://www.kaggle.com/datasets/tolgahayit/yellowrust19-yellow-rust-disease-in-wheat

1. **Written Report** (containing the following sections):
    - Abstract
    - Introduction & Problem Statement
    - Literature Review
    - Dataset Description
    - Exploratory Data Analysis
    - Visualization
    - Model Development & Evaluation
    - Results & Interpretation
    - Conclusion & Discussion

Abstract

Yellow rust is a significant fungal disease that poses a continuous threat to global wheat production, causing major yield losses for farmers. This model is focused on developing an identification model that would be able to test images and classify if it is diseased and should be removed and the area it was found monitored or completely healthy.

We trained an image identification model using the "Yellow-Rust-19" dataset from Kaggle, created by Tolga Hayit. This dataset contains a large collection of wheat images exhibiting various stages of yellow rust infection. The primary objective of our project was to create a system for detecting and assessing yellow rust disease severity through computer image analysis and machine learning algorithms. Our model takes the padded photos we uploaded and processes these photos and provides accurate diagnoses of whether or not it is infected or not, offering a practical tool for disease monitoring.

Our findings demonstrated that image-based machine learning models can serve as valuable early warning systems for farmers and even agricultural researchers. By enabling quick and accurate disease detection, this model allows for timely intervention and even more effective disease management through early detection. Potentially saving crops and reducing economic losses in wheat production.

Introduction

Wheat is an important commodity in agriculture. It is used in feeds for animals, baked goods, alcoholic beverages, and more. If wheat were wiped out due to a plant disease or drought then it would have a large impact on society. Notably, One of the most common problems that farmers face with wheat is yellow rust. Yellow rust occurs when wheat growing in cool and moist conditions is afflicted with *Puccinia striiformis* f.

sp. *tritici* (*Pst).* Yellow rust often causes a great loss of wheat yield from that farming season because it can easily spread through the fields by the release of spores into the air.

Identification of this disease is crucial in prevention. Being able to tell if the crop is diseased and must be pulled is important so that all other crops around it are not affected. Additionally being able to identify the presence of Uredia is also important as it tells you if you need to monitor the surrounding areas. The different types of classifications are made to alert the severity of the infection and how closely the other crops found around it need to be watched and observed. Best case scenario, is that we have resistant where the wheat is directly afflicted and only has visible necrosis / chlorosis and worst case is that there are spores found on the afflicted wheat

Literature Review

The data set Yellow-Rust-19 is valuable to researchers and practitioners in the field of machine learning and plant pathology. The images are meant to allow machine learning to identify and classify the severity of the wheat leaves affected by yellow rust disease. Yellow rust affects lots of wheat globally so knowing how to identify it and classify them helps people manage their crops. The files in this data set contained genetic information helpful in understanding the genetic basis of resistance to yellow rust. The phenotypic data contains observable traits displaying severity of the disease in the plant. The trait we're observing in the data set is the severity of the yellow rust disease seen in the plant.

Dataset Description

The dataset given are raw images of wheat leaves that are afflicted with yellow rust to varying degrees. From resistant, where there is visible necrosis / chlorosis but no presence of Uredia to Severe where there is barely any necrosis / chlorosis but a large amount of uredia is found. This data was gathered from wheat found in Turkey and collected from October to November of 2018. The data gathered is from multiple leaves of the same plant and these leaves were placed on an overall category of what that plant was classified as. These pictures were then grouped into the raw folder, where they were grouped on the overall health of the plant (R to S) and then processed into the yellow rust folder where the leaves were grouped by severity of the leaf itself not the plant.

Data analysis

This data set provides a valuable insight into the field of plant pathology and ai detection of rust disease in wheat crops. The purpose of this work is to inform experts in the field of a way to recognize this disease easily and how to combat a major type of disease which affects wheat crops. The machine learning model analyzed over 2000 images and was able to categorize them based on the training model. The machine is meant to be able to categorize the diverse images it is receiving and categorize them into correct groups based on their severity of yellow rust disease. The work describes how they created algorithms like deep CNN algorithm which helps identify the disease fast or the PCA algorithm which enhances image detection for yellow rust in wheat leaves. This type of training model is meant to help farmers be able to tell the severity of disease in their crop and allow them to be able to stop the spread.

Exploratory Data Analysis

The data was examined in a binary way with the pictures either belonging to Resistant where the leaf is practically healthy or severe where there is a presence of spores and necrosis on the leaf. Disease leaves showed necrosis (yellowing) and spores (small dots) on them which led to a severe classification, while fully green leaves with no yellowing of tiny spots were classified as resistant. There is also a possibility of creating a regression model instead of a binary model of classification.

Visualization

In our group project we're making an ai training machine that will detect the severity of yellow rust disease on wheat given any image of a wheat leaf. The model displays the severity and likelihood of that classification of yellow rust disease. This model is meant to help a worker in this field be able to identify the problem of this disease early and treat it fast. Additionally the images taken were padded with code due to the different variation in sizes causing an error when it came to the model reading the image as it reads only a specific area of the leaf and most of the time it is just the middle part of the leaf usually looking undiseased.

Model development and evaluation

A basic overview of model development was while developing the model we attempted to use a google teachable machine to train and identify the disease. At first the model was not working and we thought initially it was due to the difference in image sizes which led us to pad the images. After that the model was still not working properly thus we had to edit around the code more, however the model eventually became unworkable and we found certain discrepancies where in google teachable machine the identification process was coming out good but when it comes to copilot it was not identifying the leaf properly at all. We then tried to train the model with python code instead of using the teachable machine but the similarities between the categories caused issues with loading and training the model. This led us to keeping the same concept of classifying various different images into their proper disease category but instead of a regression model we went with a binary model. With this new concept we were able to use the google teachable machine to train model and it was able to accurately identify the image categories.

Some important things to note about the modes is that it went through multiple edits and troubleshoots when creating the teachable machine. We initially built a model to identify 6 different classification of severity of yellow-rust within wheat. However due to some errors, the model was cut down to just(R) and (S) which are Resistant and Severe. Now this code is able to identify whether there is any yellow-rust present in general, but doesn't differentiate based upon severity of the disease. The coding for the teachable machine was done using around 10,000+ pictures at first but these caused the machine to lag greatly due to the sheer volume of pictures thus it was cut down to 4,000. These images used were also needed to be padded beforehand, as when it was not the result accuracy was 50% which is as good as a blind guess. In colab after all the following was done resulted in a 86% accuracy rate regarding a correct reading of R or S.

Results and Interpretation

After multiple troubleshooting, the code ended up running on a binary basis determining if the leaf is either not diseased (aka Resistant) or diseased (aka Severe). The initial goal for this model was to use it as a regression model however multiple problems arose due to how close the severity of disease scale. At first the machine was only identifying everything as no disease. This was due to a combination of things, such as, the machine only reading a specific chunk of the image and the images being of different sizes. This led us to writing a code to pad all the images to make them all be the same size. After that we still ran into issues with a discrepancy between the teachable machine and exporting the code as there for some reason was a change throughout the process which led to images being misidentified. At the end we decided to use a binary model that tells us yes or no the leaf is diseased or not diseased. After tweaking the code and using the padded images, we yielded results with an identification accuracy rating of 86% instead of the initial 50% without the padded image. This is statistically better as a 50% is the same as a blind guess while an 86% signifies that there is an element of certainty. The padded images clearly have a large impact on the training of the model so it should be noted that size similarity of the testing and training images are important.

Conclusion and Discussion

This model is a good starting point for leaf disease classification. The dataset is helpful especially in the training of the model. Although our model ended up working there is still much left to be done. The biggest challenge is the fact that all images need to be of the same size before it can be run properly and result in a high accuracy rate. The training model also needs to be able to distinguish between regression pictures as if pictures are too similar the model tends to get confused and mess with the training data. In the future this model can be converted into an app that farmers can use to identify whether or not their plants are disease and hopefully how diseased the plant is and if there is a need to monitor a certain circumference. It can also be incorporated hopefully into security cameras and be used to monitor the crops more easily and that farmers wont have to keep going out and checking around their large fields.