Overview

A set of images of normal seeds are provided. In 24 hours, some of these seeds start to become purple eater monsters, that is they develop purple coloration. Images of the seeds after 24 hours are also provided.

About 173 pairs of Images of seeds taken on day 0 and 24 hours later are provided. The Day 0 images are of normal seeds. The Day 1 images shows seeds after 24 hours. Some seeds are normal while some seeds have developed a purple tint.

The purpose of this challenge is to use these images and propose solutions to predict the seeds that could turn purple.

Methods

The proposed solution has been developed by analysing the available information and experimental testing on the seed images provided.

Online documentation on Image processing and machine learning libraries such as OpenCv (https://opencv.org) and scikit learn (https://scikit-learn.org) have been referred to.

Materials

The proposed solution is developed in Python 3.7. The dependent modules are –

- Numpy
- Pandas
- OpenCV
- Matplotlib
- Scikit Learn

Discussion

The proposed solution for this challenge involves using digital image analysis and machine learning techniques to develop the required algorithm.

Image analysis techniques are used to extract morphological (seed shape) information (such as length, width, area, etc) and truth values (turned purple or not).

As the requirement is to predict if the seeds turn purple or not, it is a binary classification problem. Suitable binary classification algorithms are applied on the seeds data extracted and applied to unseen/test seeds images for prediction.

Steps to develop the prediction algorithm

1. Seeds Data preparation
   Image Analysis and extraction of seeds information
   Feature Engineering
   Data Preparation for modelling
2. Develop prediction model and train on training data
3. Fit model to test data, make predictions and summarize results

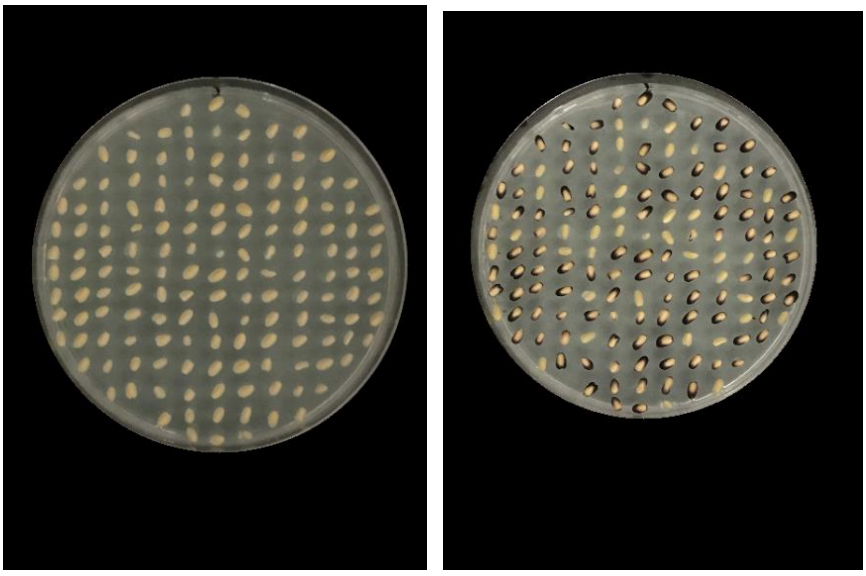
Seed Data Preparation

1. Image Analysis

   Step 1: Crop the outer circle of the image

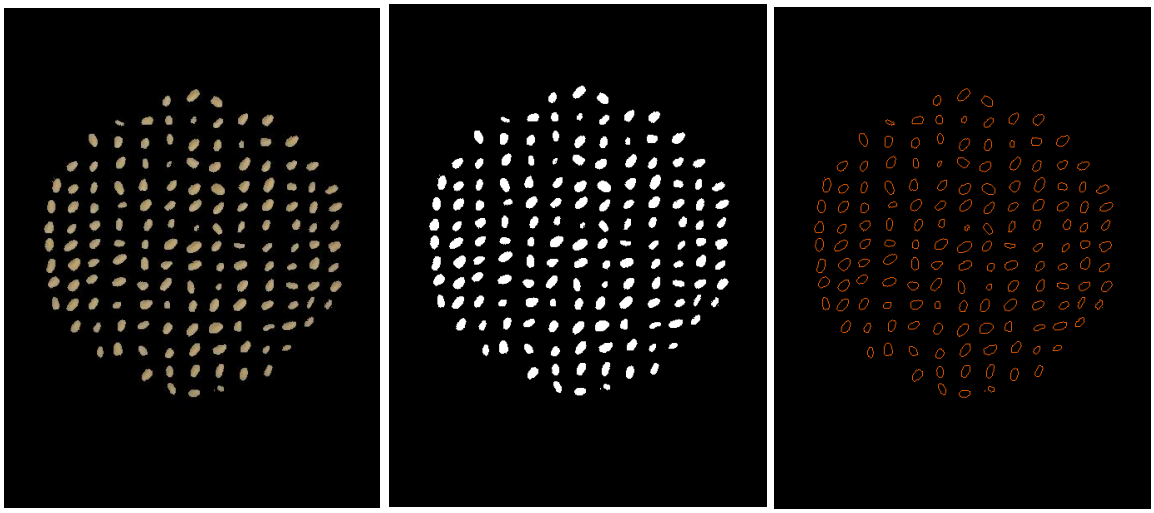
   Each Day 0 image is read. The outer circle of the image is identified using cv2 functions Canny edge detection and HoughCircles. This circle is imposed on the read image and the circle is cropped out. Similarly the outer circle of Day 1 image is cropped.

   The cropped images would look like

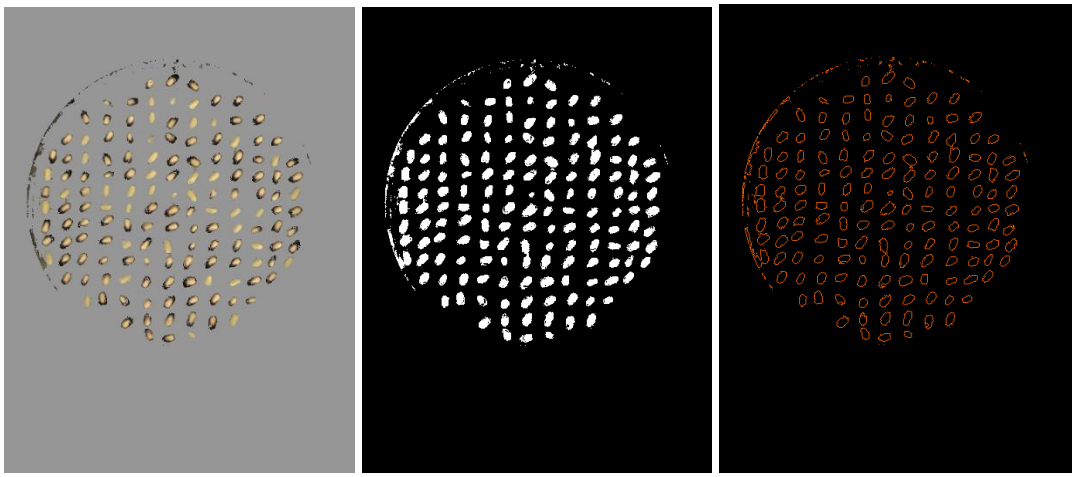


Step 2: Segment seeds and identify contours of seeds

The Day 0 image is converted to HSV format. A mask is created in the HSV image for the color range corresponding to the color of seeds. This would mark out areas of seeds in white with black background. From this the contours of the seeds are extracted.



Similarly, the Day 1 image is converted into HSV format. This image has normal seeds as well as the seeds with purple tint. Hence the mask is created for the normal seed color and purple color ranges. Separate out the seed areas and extract the contours of the seeds.



Step 3: Extract seed information and prepare seeds dataset

For each contour of seed in Day 0 image –

- From Image name, obtain seed variety (DHnnn)
- Number the seed
- Obtain the bounding rectangle of the seed contour – Expressed as x,y co-ordinate of left most corner, width and height of rectangle
- Obtain Area and Perimeter of the seed contour

- Extract Texture information – Laws Texture Energy Measures
- In the corresponding seed contour of Day 1 seed contour, check if purple coloration is present – If present, set truth value to 1, else set to 0

The above extracted information is consolidated into seeds dataset.

2. Feature Engineering

Compute the following new features and add to seed dataset –

- Circularity = 4 * Pi * Seed Area / Perimeter ^ 2
- Aspect Ratio = Width / Length

3. Data Preparation for modelling

- Apply label-encodings to the following fields
  - Seed Variety
- Normalize the following fields
  - Length
  - Width
  - Area
  - Perimeter
  - Texture Score
- Split the data into training (80% of data) and validation (20% of data) sets

Develop Prediction Model

As this is a supervised binary classification problem, apply the following machine learning models on training dataset.

- Logistic Regression
- Decision Tree Classifier
- Support Vector Machine
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Random Forest Classifier
- K-Nearest Neighbors
- Naive Bayes

The performance of the models can be measured using F1 score.

F1 Score = 2 * [(Precision * Recall) / (Precision + Recall)]

Precision = True Positive / (True Positive + False Positive)
Recall = True Positive / (True Positive + False Negative)

Compute feature importance and discard irrelevant features. Rerun the models and compute revised metrics scores. Select the model with the best score.

Make Predictions and Summarize Results

Extract the seed information from test images and prepare test dataset as described in 'Seed Data Preparation' section.
Fit the selected model on test data and predict the outcome on seeds in test images.
Draw a bounding box around the seeds that have been predicted as turning purple.
Create an output csv file consisting of Image ID, Seed Number and Predicted Outcome.

Data

About 173 pairs of Images of seeds taken on day 0 and 24 hours later are provided. From these images, the seeds data is extracted. Additional ratios are computed (as described in Feature Engineering section).

As the proposed solution is largely based on image analysis, improving the clarity and contrast between the different regions of the image would improve outcome. For example, in Day 1 images, the purple coloration of the seeds is very close to the seed container border (dark colored outer circle). Changing the container border to a color that is different from the seed and purple colors will improve seed segmentation.

The following additional data related to seed and environmental conditions could be provided. These could impact the seed growth and hence could improve the algorithms performance.

- Temperature
- Humidity
- Seed grade – If any classification (such as good, average, poor, etc) is available
- Any other information that impacts the seed growth.

Assumptions/Risks

This proposed solution is largely dependent on the seed shape information extracted from the images provided. Hence the images need to be carefully and consistently photographed.

It is assumed that -

1. The seed positions are not changed between Day 0 and Day 1 images.
2. The same photographic settings are used to take the images. The camera position, resolution, image size and illumination are consistent for all images.

Expected Results

The proposed solution is expected to provide 2 outputs –

- Output csv file containing predicted outcomes
- Test images with bounding box drawn around seeds that are predicted to turn purple.