

1 Introduction

- Inari is a biotechnology company improving crops using gene-editing and AI
- Did Inari's genetic modifications make a difference in the physical appearance of soybean plants?
- Physical differences between genetically modified and unedited plants are difficult for humans to interpret. We used machine learning on ~1,000 images to test whether computers can distinguish physical features

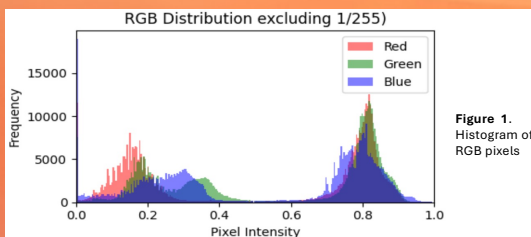
2 Inari's Datasets



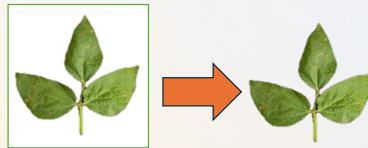
We focused on leaf trifoliolate due to the large dataset size

3 Exploratory Data Analysis (EDA)

- Histogram of pixel colors (Fig. 1)
- Image dimensions
- Number of images per class
- Vegetation indices (greenness measurements)



4 Images and Segmentation



EDA showed peak at green pixels, used color to segment. Methods:

- Masking
- OTSU
- SAM

5 Feature Extraction

Not all ML models use image input, we manually calculated metrics (features), creating tabular data:

- RGB Values
- Contour length (Fig. 2)
- Leaf Area
- Leaf shape metrics
- DGCI
- MPRI
- HSV Mean and Standard Deviation
- Centroid

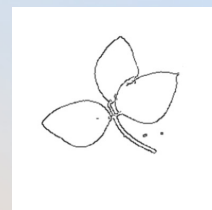
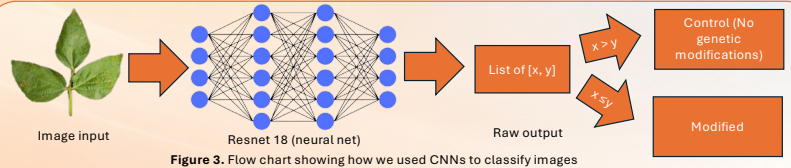


Figure 2. Example of feature extraction. We extracted contour of leaves and then calculated its length



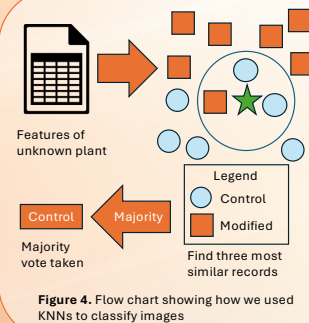
6 ML Models

Convolutional Neural Nets (uses image input):

- Visual Geometry Group 19
- Resnet18
- Custom made neural nets x 3

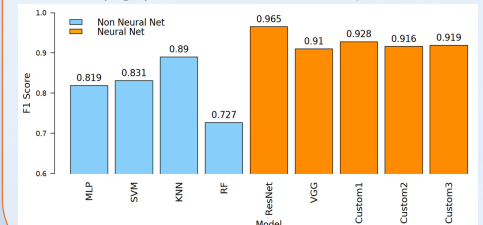
Non Neural Nets (uses tabular input):

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Multi-Layer Perceptron (MLP)



7 Results

- Best neural net was Resnet18 (Fig 3), with ~96% accuracy
- Best non neural net based approach was KNN (fig4), with ~92% accuracy



8 Conclusions

- Our binary classifier models (CNNs and KNN) reliably classified leaves as genetically modified or not based on evaluation metrics
- Modifications made a systematic difference in the physical appearance of plants, even though this difference may not be interpretable to humans

9 Future Goals

- Train models that can distinguish specific genotypes with high accuracy
- Use clustering methods to group the data and identify features that differentiate between genotypes
- Explore additional aspects, such as roots and whole plant images

Acknowledgements

We would like to thank Gustavo Santiago (TA), Dylan Elliot, Matt Eckerle, Garrett Bernstein, Ross Altman (Inari Team), and The Data Mine for making this possible