# Plant Seedling Classification Using CNN

Plant Seedling Classification Project – PGP ML/AI

July 23, 2023

# Contents / Agenda

# Executive Summary

- The goal of the project is to create a classifier capable of determining a plant's species from an image to reduce expense, manual labor, time, and inefficiency and to improve crop yields, free-up time for higher-order decision making, and promote sustainability
- A dataset containing 4,750 images of 12 unique plant species was used to train and test models
- An exploratory data analysis was completed, the images/labels were preprocessed, and data was split into 80% training/10% validation/10% test sets
- Two convolutional neural network model architectures were built, trained, and tested, and performance on evaluation metrics were evaluated
- The final chosen model is able to classify plant seedlings into their respective categories at an accuracy of 77%

# Business Problem Overview and Solution Approach

## Problem

- Manually sorting and recognizing different plants and weeds is expensive, labor intensive, time consuming, and inefficient

## Solution

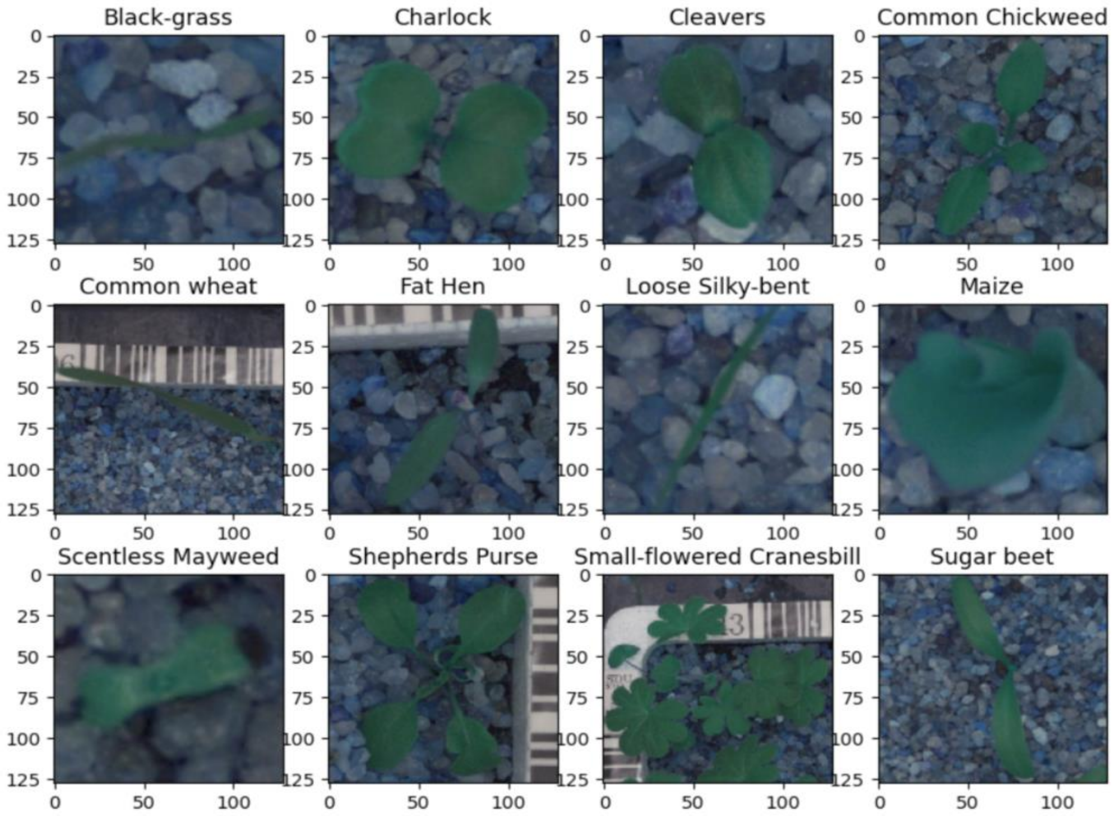
- Use machine learning to classify plant seedlings more efficiently and effectively which can lead to better crop yields, free up time for higher-order agricultural decision making, and result in more sustainable environmental practices

## Methodology

- Step 1: Preprocess dataset containing images of unique plant species
- Step 2: Train Convolutional Neural Network Models
- Step 3: Evaluate and compare results and choose final model

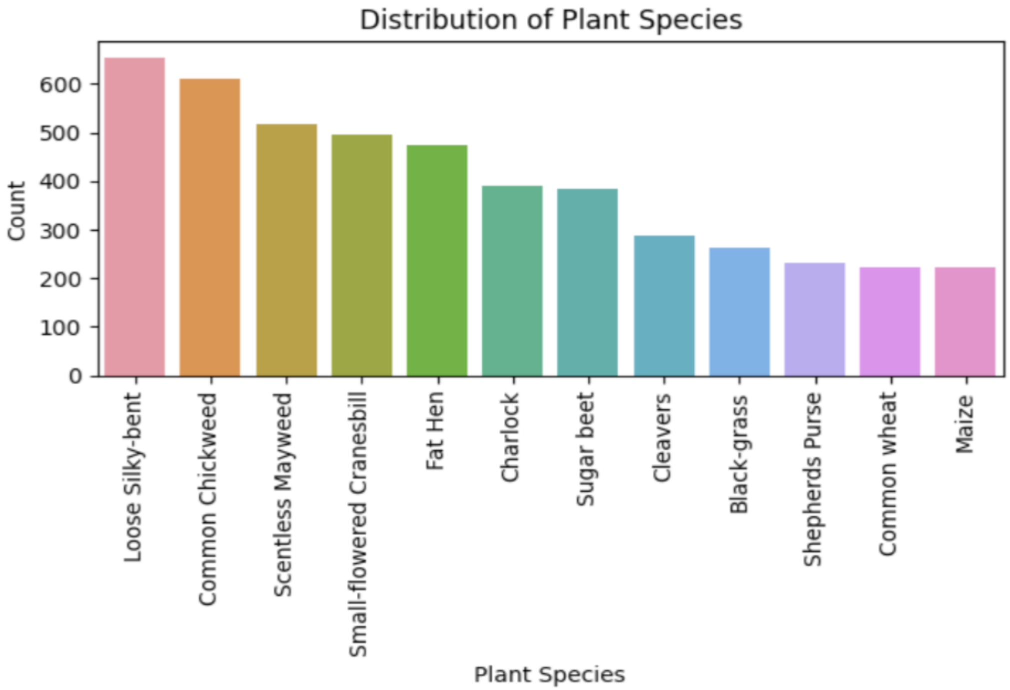
# EDA Results

- Overview of data:
    - 4,750 color images
    - 128x128 pixels each
    - 3 color dimensions (BGR)
    - 12 unique plant species
    - Separate images and labels files



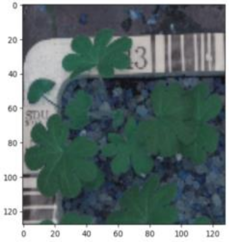
# EDA Results (contd.)

- Imbalanced dataset
  - Ranges from 654 images (Loose Silky-Bent) to 221 images (Common Wheat, Maize)
- Data augmentation can help with imbalanced datasets

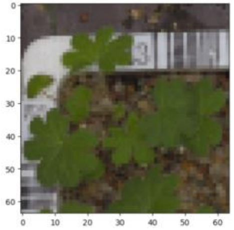


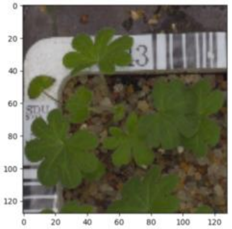
# Data Preprocessing

Original image



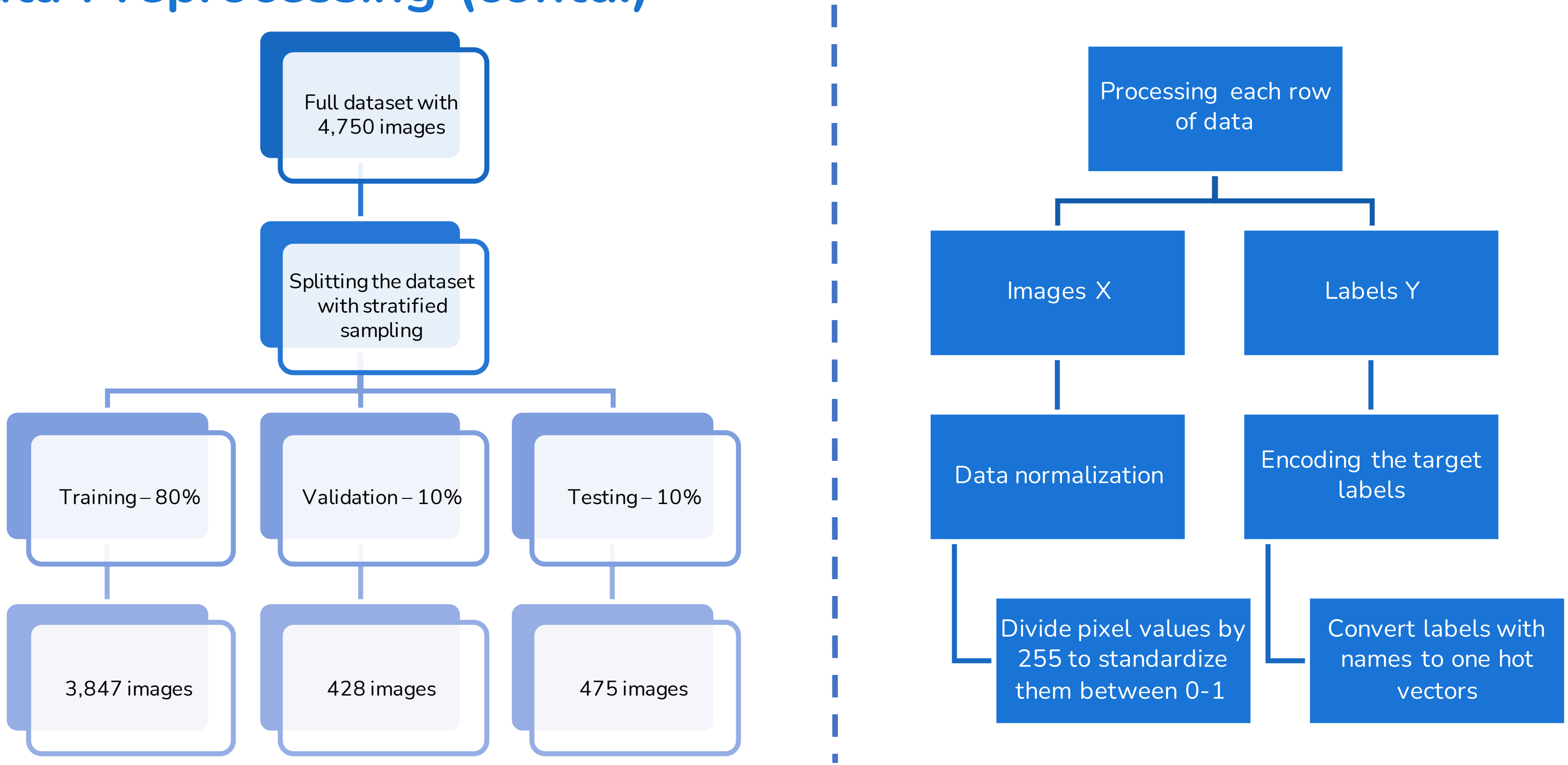
Resized images from 128x128 to 64x64 pixels to reduce computational expense





Converted BGR color format to RGB color format

# Data Preprocessing (contd.)



# Model Architecture

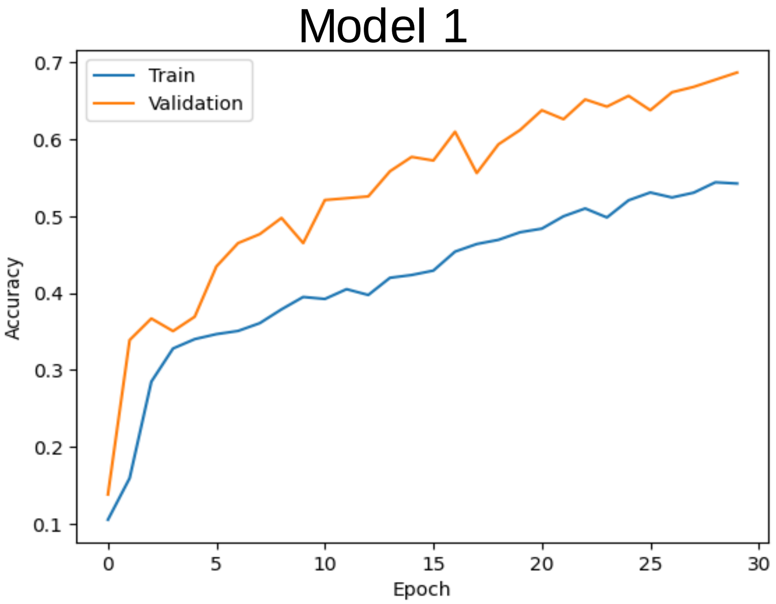
- All models utilized:
  - Convolution layers: padding = ’same’, activation = ’relu’, 3x3 kernel size
  - Pooling layers: Max pooling, padding=’same’, 2x2 patches
- Model 1:
  - Adam optimizer, epochs = 30, batch size = 32
  - 128,828 parameters

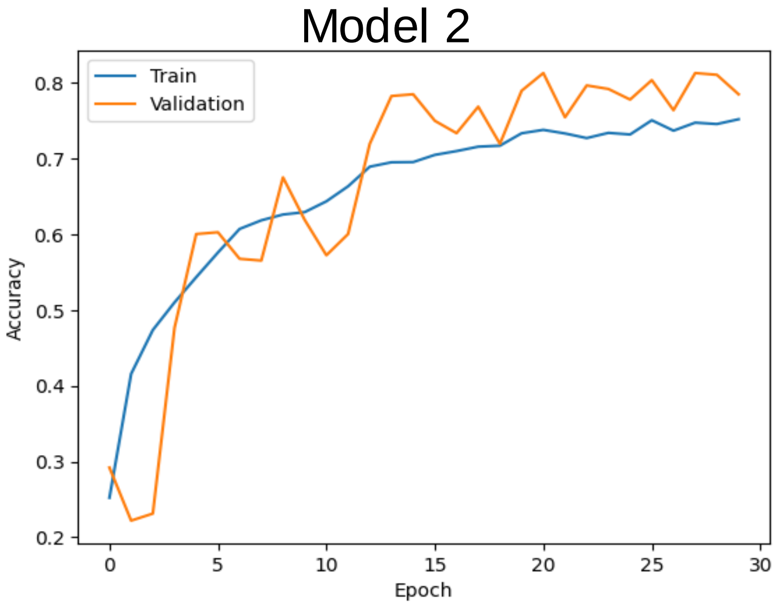
# Model Architecture (contd.)

- Model 2:
  - Applied data augmentation by randomly rotating images 20 degrees
  - Applied function to decrease learning rate by a 0.5 factor if loss is not decreasing
  - Adam optimizer, epochs = 30, batch size = 64
  - 151,676 parameters

# Model Performance Summary

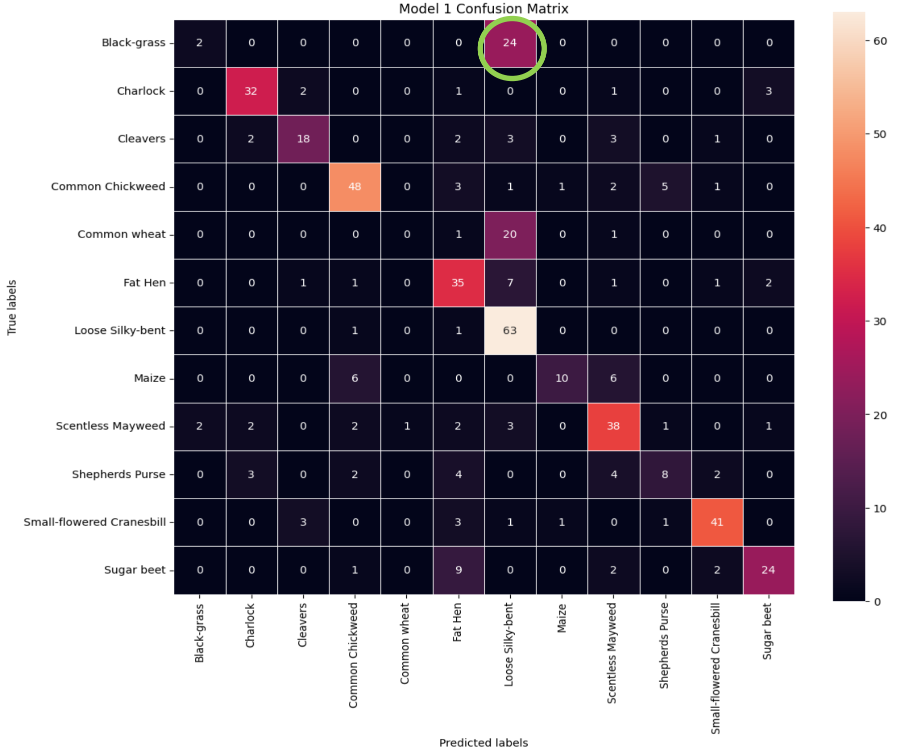
- Model 1 was underfitting as the validation did better than the training set
- Under/Overfitting reduced in Model 2 with Batch Norm, reduced learning rate, and data augmentation

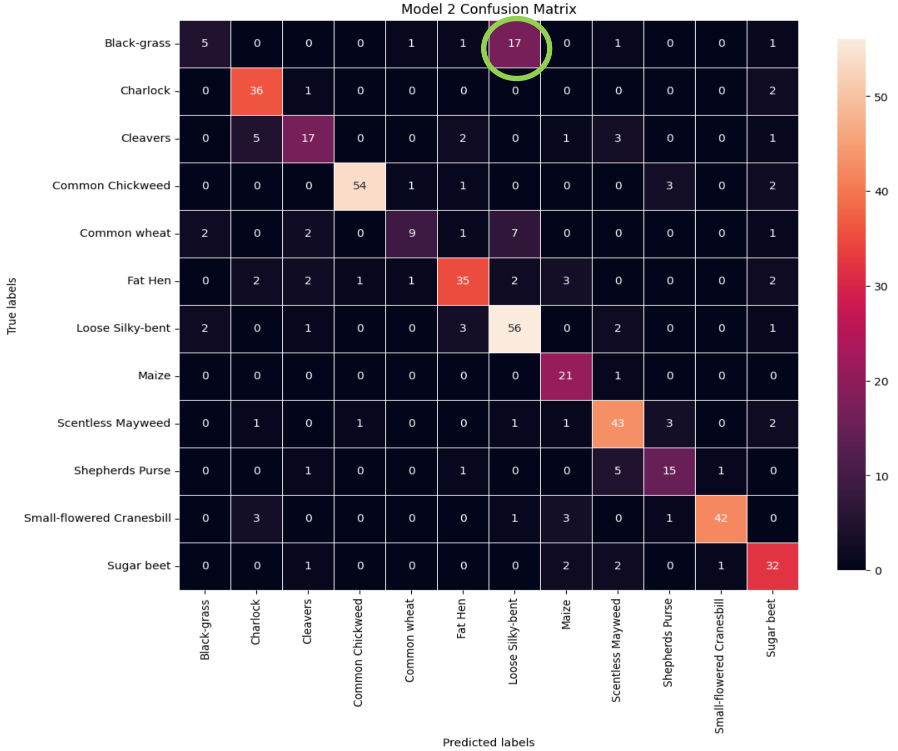


# Model Performance Summary (contd.)

- Numbers off the main diagonal are misclassifications
- Black-grass was most commonly misclassified among both models as Loose Silky-bent



# Model Performance Summary (contd.)

| Metric | Model 1 | Model 2 | Improvement |
|---|---|---|---|
| Training Accuracy | 54% | 75% | +21% |
| Validation Accuracy | 69% | 79% | +10% |
| Testing Accuracy | 67% | 77% | +10% |
| Precision (weighted avg) | 66% | 77% | +11% |
| Recall (weighted avg) | 67% | 77% | +10% |
| F1 Score (weighted avg) | 64% | 76% | +12% |

# Conclusion

- Model 2 outperformed Model 1 in every metric and showed less underfitting and overfitting
- Model 2 can classify plant seedlings with 77% accuracy
- Model 2 utilized data augmentation, decreased learning rate, batch normalization, and higher batch size than Model 1 to improve performance
- Recommendations to improve performance:
    - Implement transfer learning
    - Train model with grayscale images or masking the background
    - Focus on Loose Silky-bent and Black-grass differences as that was commonly misclassified
    - Apply other data augmentation techniques like zooming in and flipping
    - Add spatial dropout and more batch norm layers to further reduce overfitting
    - Increase the number of fully connected layers

# APPENDIX

# Data Background and Contents

- The Aarhus University Signal Processing group, in collaboration with the University of Southern Denmark, released a dataset containing images of unique plants belonging to 12 different species.
- Data provided in two separate files:
  - Images.npy
  - Label.csv
- List of Species:
  - Black-grass
  - Charlock
  - Cleavers
  - Common Chickweed
  - Common Wheat
  - Fat Hen
  - Loose Silky-bent
  - Maize
  - Scentless Mayweed
  - Shepherds Purse
  - Small-flowered Cranesbill
  - Sugar beet

# Happy Learning !