Plant Seedlings Classification

Project #5- Introduction to Computer Vision

Tuesday, November 12, 2024

Contents / Agenda

Executive Summary

Executive Summary

The agricultural industry, traditionally reliant on manual labor for plant growth monitoring, faces inefficiencies due to the time-consuming process of sorting and identifying plants. Despite advancements in technology, the need for human intervention in plant identification persists. This project seeks to modernize the sector by leveraging AI and Deep Learning to automate plant seedling identification, significantly reducing labor costs and improving speed and accuracy. By building a Convolutional Neural Network (CNN) to classify images of 12 different plant species, the project aims to outperform manual methods, enhance crop yields, and allow workers to focus on more strategic agricultural tasks.

Executive Summary- Business Recommendations

After evaluating the performances of each of the two CNN models designed to identify and classify plant seedlings into their respective categories, the following observations and business recommendations can be made as follows:

1. **Enhanced Data Collection for Underperforming Classes:** To improve model performance, particularly for the species with the lowest prediction rates, ***invest in collecting more comprehensive and varied data samples for these species***. This could help the models learn more robust features about these plants.

2. **Feature Engineering and Augmentation:** Given the consistent performance in certain classes, it may be beneficial to ***analyze the features that led to high accuracy and see if similar features can be engineered for the underperforming classes.*** Additionally, using techniques like ***image augmentation*** could help improve the model's ability to generalize across less common species.

3. **Focused Research on Hard-to-Classify Species:** ***Deploy targeted research to understand why certain species are harder to classify.*** This might include botanical studies to find distinct visual markers not currently captured in the dataset.

Executive Summary- Business Recommendations

4. **Iterative Model Improvement: *Use an iterative approach to model training*** where insights from current model performance directly inform subsequent rounds of model refinement and data collection strategies.

5. **Deployment in Staged Environments: *For deployment, consider using the models in a staged approach*** where they first handle identification tasks they perform well in and gradually expand to more complex classifications as model confidence and performance improve.

6. **Integration with Agricultural Decision Systems: *Integrate these models into larger agricultural decision-making frameworks*** where AI-driven identifications can help optimize pesticide use, planting strategies, and crop rotation plans, thus enhancing yield efficiency and sustainability.

By implementing these strategies, the agricultural sector can significantly benefit from reduced labor costs, improved accuracy in plant and weed management, and enhanced overall crop management strategies, leading to better yields and reduced environmental impact.

Business Problem Overview

&

Solution Approach

Business Problem Overview

The agricultural industry has long relied on manual labor to monitor and assess plant growth, often requiring workers to spend significant time sorting and identifying different plants and weeds. Despite advancements in agricultural technology, the need for human intervention in plant identification remains extensive, leading to inefficiencies and a significant allocation of time and effort. This situation presents a clear opportunity for modernization, as the agriculture sector, which is valued in the trillions, has the potential to be transformed by cutting-edge technological innovations.

One of the key areas ripe for disruption is the identification of plant seedlings, which remains a labor-intensive task. Artificial Intelligence (AI) and Deep Learning can play a crucial role in addressing this challenge. By leveraging AI to automate the identification process, the time and energy spent on sorting plant seedlings can be significantly reduced. Furthermore, **AI has the potential to outperform human labor in terms of speed and accuracy, leading to better crop yields and enabling workers to focus on higher-order tasks, such as strategic agricultural decision-making.**

Solution Approach

This project aims to build a CNN to classify plant seedlings, utilizing a dataset that includes images of 12 different plant species. The goal is to create a model that can accurately classify plant species from images, enabling faster, more accurate identification than manual methods.

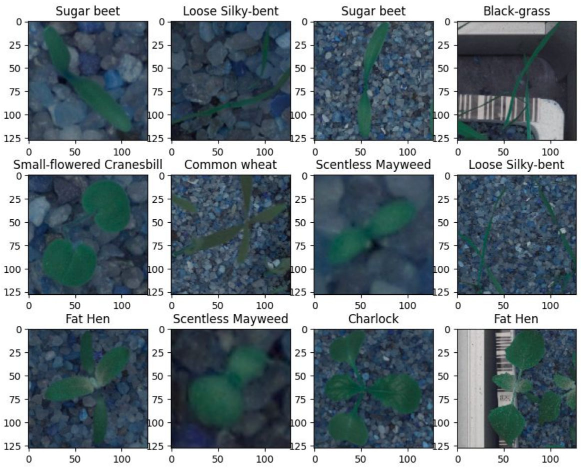
The project's methodology includes data preprocessing steps like resizing images for computational efficiency, splitting data into training, validation, and testing sets, and using a one-hot encoding approach for label representation. The CNN model is then trained, with evaluation metrics including accuracy, recall, and F1-score, followed by strategies for model performance improvement such as learning rate reduction and data augmentation.

Exploratory Data Analysis

EDA Results

Images: (4750, 128, 128, 3)
Labels: (4750, 1)

Shape of Dataset



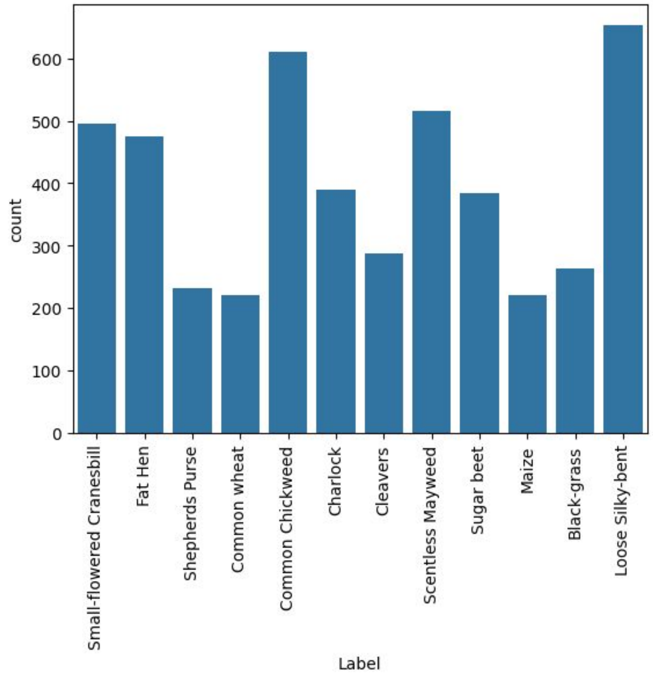
Plotting random images from each class

Input the images and labels and plot the images with their labels.

Subplot rows = 3

Subplot columns = 4

Figure size: 10 x 8



Checking the distribution of target variable

Check for data imbalance.

EDA Results

It can be observed that the distribution of seedling samples in the training dataset exhibits a significant imbalance, with the number of samples ranging from over 600 for the most represented species to just 220 for the least represented. This discrepancy can lead to poorer predictability for under-represented classes, as the model has fewer examples to learn from compared to classes with a larger number of samples.

Seedling counts:

1. Loose Silky-bent (600+)
2. Common Chickweed (600)
3. Scentless Mayweed (540)
4. Small-flowered Cranesbill (500)
5. Fat Hen (480)
6. Charlock (400)
7. Sugar Beet (400)
8. Cleavers (290)
9. Black-grass (275)
10. Shepherds Purse (230)
11. Common Wheat (220)
12. Maize (220)

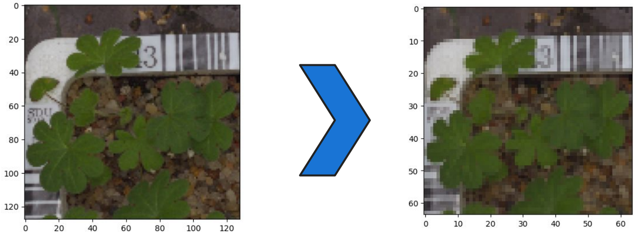
Data Preprocessing

Data Preprocessing

- As we have less images in our dataset, we will only use 10% of our data for testing. The remainder will be split 10% into our data for validation and 80% into our data for training.

- **Resizing Images**

  Reduced image size from 128 x 128 to 64 x 64 pixels



- **Encoding the target labels**

  Convert labels from names to one hot vectors using *LabelBinarizer*

  y_train = enc.fit.transform (3847 , 12)

  y_val = enc.transform (428 , 12)

  y_test= enc.transform (475 , 12)

- **Data Normalization**

  Normalization will be scaling method used; all pixel values are divided 255 to standardize the images to have values between 0 - 1.

15/15 - 1s - 47ms/step - accuracy: 0.6147 - loss: 1.1419

Model Performance Summary

Model Performance Summary- Model 1

Convolutional Layers (Conv2D) = 3
Pooling Layers (MaxPooling2D) = 3
Flatten Layers (Flatten) = 1
Dense Layers (Dense) = 2
Dropout Layers (Dropout) = 1

Layers = 10

Total parameters: 128,828
Trainable parameters: 128,828
Non-trainable parameters: 0

Observations:

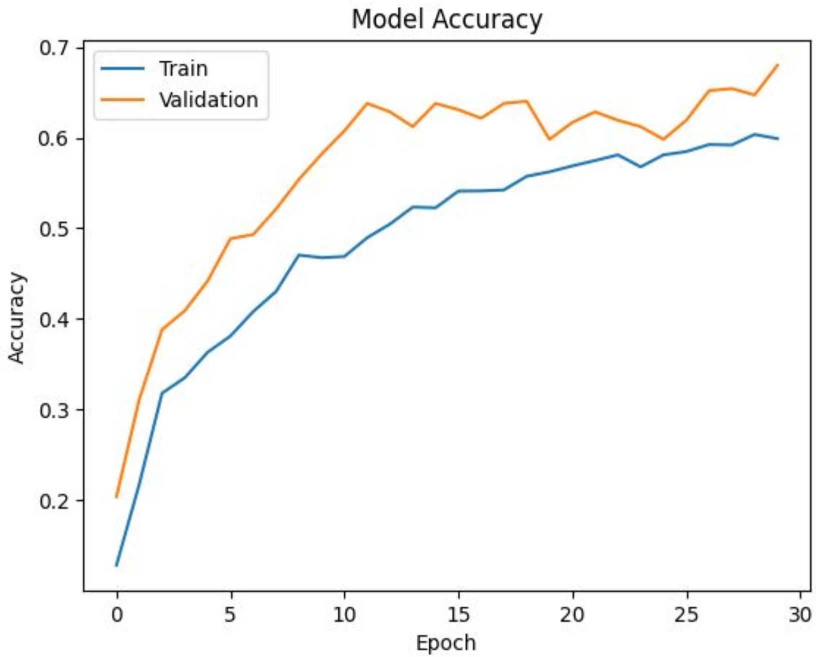
- *10 layers*
- *Perfect non-trainable parameter score of 0.*
- *Indicates an efficient model configuration with full optimization potential. The absence of non-trainable parameters suggests that all components of the model contribute directly to the learning process, potentially improving performance and adaptability.*

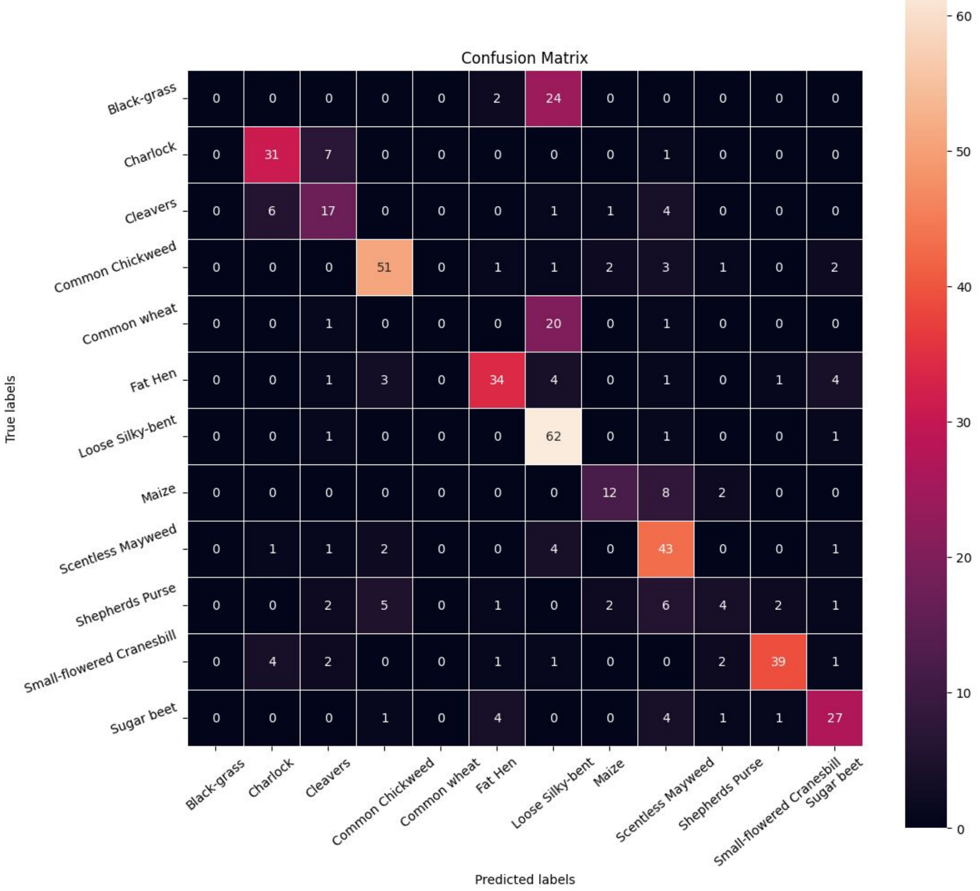
Model: "sequential_4"

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 64, 64, 128)	3,584
max_pooling2d_12 (MaxPooling2D)	(None, 32, 32, 128)	0
conv2d_13 (Conv2D)	(None, 32, 32, 64)	73,792
max_pooling2d_13 (MaxPooling2D)	(None, 16, 16, 64)	0
conv2d_14 (Conv2D)	(None, 16, 16, 32)	18,464
max_pooling2d_14 (MaxPooling2D)	(None, 8, 8, 32)	0
flatten_4 (Flatten)	(None, 2048)	0
dense_8 (Dense)	(None, 16)	32,784
dropout_4 (Dropout)	(None, 16)	0
dense_9 (Dense)	(None, 12)	204

Total params: 128,828 (503.23 KB)
Trainable params: 128,828 (503.23 KB)
Non-trainable params: 0 (0.00 B)

Model Performance Summary- Model 1





Top 3 correctly predicted labels:

1. Loose-silky bent (62)
2. Common Chickweed (51)
3. Small-flowered Cranesbill (39)

Bottom 3 correctly predicted labels:

1. Black Grass (0)
2. Common Wheat (0)
3. Shepherds Purse (4)

Model Performance Summary- Model 1

The model achieved an overall accuracy of 67% on the test data, with a loss of 1.05. Key metrics such as precision, recall, and F1-score varied across different classes, indicating that the model performed better on some classes than others. Specifically:

- **Best Performance:** Classes like 1, 3, 5, 6, 8, 10, and 11 demonstrated relatively good performance, with high recall values, especially class 6 with a recall of 0.95, indicating the model was effective in correctly identifying instances of these classes.
- **Poor Performance:** Classes 0, 4, and 9 had low precision, recall, and F1-scores, with 0 and 4 being particularly poor as they had zero recall and precision, indicating that the model was unable to correctly identify instances for these classes.
- **Macro and Weighted Average:**
  - **Macro Avg:** Precision = 0.56, Recall = 0.58, F1-Score = 0.56, showing the overall performance across classes is relatively low, largely affected by the poor prediction of some minority classes.
  - **Weighted Avg:** Precision = 0.62, Recall = 0.67, F1-Score = 0.63, indicating a slightly better performance when considering the support for each class, but there is still room for improvement.

Overall, while the model was fairly accurate in identifying some classes, there is a notable imbalance, especially in its inability to correctly identify certain classes (like 0 and 4). This could suggest that the model might be overfitting or undertrained for certain underrepresented classes in the dataset.

Classification Report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	26
1	0.74	0.79	0.77	39
2	0.53	0.59	0.56	29
3	0.82	0.84	0.83	61
4	0.00	0.00	0.00	22
5	0.79	0.71	0.75	48
6	0.53	0.95	0.68	65
7	0.71	0.55	0.62	22
8	0.60	0.83	0.69	52
9	0.40	0.17	0.24	23
10	0.91	0.78	0.84	50
11	0.73	0.71	0.72	38
accuracy			0.67	475
macro avg	0.56	0.58	0.56	475
weighted avg	0.62	0.67	0.63	475

Evaluation on Test Data

```
15/15 - 1s - 79ms/step - accuracy: 0.6737 - loss: 1.0498
```

Model Performance Summary- Model 2

Convolutional Layers (Conv2D) = 2
Pooling Layers (MaxPooling2D) = 2
Batch Normalization Layers (BatchNormalization) = 1
Flatten Layers (Flatten) = 1
Dense Layers (Dense) = 2
Dropout Layers (Dropout) = 1

Layers = 9

Total parameters: 151,676
Trainable parameters: 151,612
Non-trainable parameters: 64

Observations:

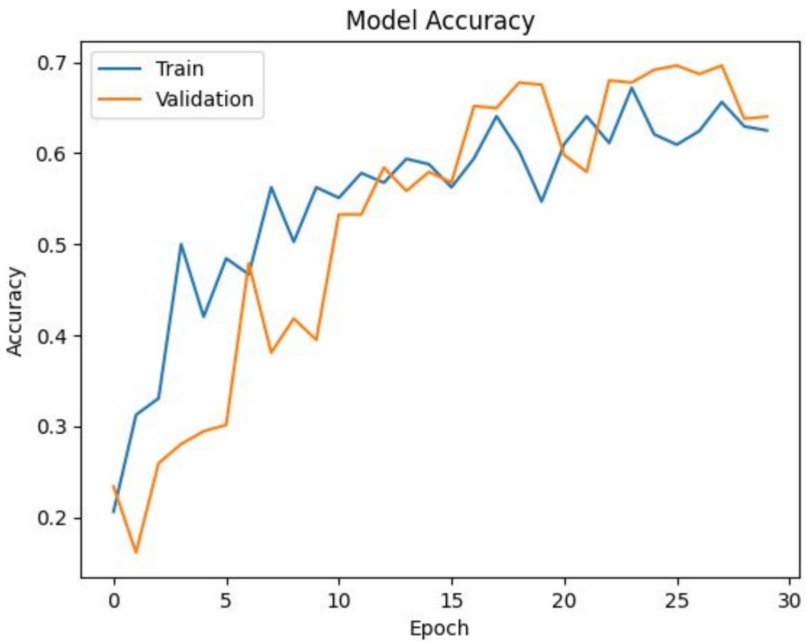
- *9 layers*
- *Non-trainable parameter score of 64 could indicate that certain layers, like the batch normalization layer, have fixed parameters that do not update during training. This architecture offers a balance of complexity and regularization through dropout and batch normalization, aimed at improving generalization and model stability.*

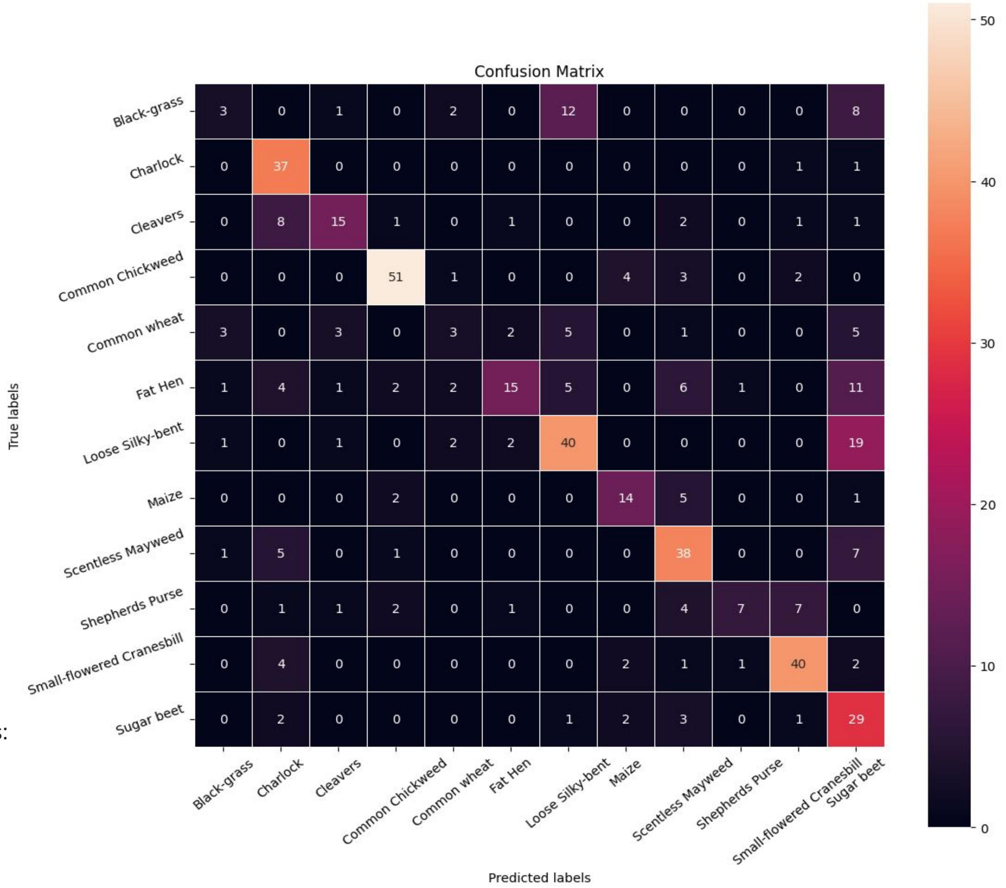
Model: "sequential_3"

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 64, 64, 64)	1,792
max_pooling2d_2 (MaxPooling2D)	(None, 32, 32, 64)	0
conv2d_3 (Conv2D)	(None, 32, 32, 32)	18,464
max_pooling2d_3 (MaxPooling2D)	(None, 16, 16, 32)	0
batch_normalization_1 (BatchNormalization)	(None, 16, 16, 32)	128
flatten_1 (Flatten)	(None, 8192)	0
dense_2 (Dense)	(None, 16)	131,088
dropout_1 (Dropout)	(None, 16)	0
dense_3 (Dense)	(None, 12)	204

Total params: 151,676 (592.48 KB)
Trainable params: 151,612 (592.23 KB)
Non-trainable params: 64 (256.00 B)

Model Performance Summary- Model 2





Top 3 correctly predicted labels:

1. Common Chickweed (51)
2. Loose-silky bent (40)
3. Small-flowered Cranesbill (40)

Bottom 3 correctly predicted labels:

1. Black Grass (3)
2. Common Wheat (3)
3. Shepherds Purse (7)

Model Performance Summary- Model 2

The model achieved an accuracy of 61% on the test data, with a loss of 1.14. The performance varies significantly across different classes:

- **Best Performance:** Classes such as 3, 1, 10 performed well with high recall, precision, and F1-scores, indicating the model was able to identify these classes accurately. Class 3 performed particularly well with 0.86 precision and 0.84 recall.
- **Moderate Performance:** Classes like 6, 7, 8, 9 performed decently, with balanced precision and recall scores. For example, class 6 has a 0.63 precision and 0.62 recall, showing relatively balanced performance for the class.
- **Poor Performance:** Classes 0, 4, 5, and 11 have relatively low precision, recall, and F1-scores. Class 0 had 0.33 precision and 0.12 recall, meaning the model struggled significantly to identify this class, and 4 and 5 had low overall performance as well.
- **Macro and Weighted Average:**
  - **Macro Avg:** Precision = 0.61, Recall = 0.56, F1-Score = 0.55 — shows the model's overall performance across classes is moderate, but it is heavily influenced by the poor performance of some underrepresented classes.
  - **Weighted Avg:** Precision = 0.64, Recall = 0.61, F1-Score = 0.60 — slightly better when taking into account the class support, but still points to room for improvement, particularly for underperforming classes.

The model performed reasonably well for some classes but struggled with others, particularly underrepresented classes. The low recall and precision for certain classes suggest that the model may need additional training, rebalancing (e.g., through data augmentation or oversampling), or adjustments to improve performance across all categories. Further refinement is needed to handle the imbalanced class distribution and ensure more consistent performance across all classes.

Classification Report

	precision	recall	f1-score	support
0	0.33	0.12	0.17	26
1	0.61	0.95	0.74	39
2	0.68	0.52	0.59	29
3	0.86	0.84	0.85	61
4	0.30	0.14	0.19	22
5	0.71	0.31	0.43	48
6	0.63	0.62	0.62	65
7	0.64	0.64	0.64	22
8	0.60	0.73	0.66	52
9	0.78	0.30	0.44	23
10	0.77	0.80	0.78	50
11	0.35	0.76	0.48	38
accuracy			0.61	475
macro avg	0.61	0.56	0.55	475
weighted avg	0.64	0.61	0.60	475

Evaluation on Test Data

```
15/15 - 1s - 47ms/step - accuracy: 0.6147 - loss: 1.1419
```

Model Performance Summary- Model 1 vs Model 2

Model 1 Classification Report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	26
1	0.74	0.79	0.77	39
2	0.53	0.59	0.56	29
3	0.82	0.84	0.83	61
4	0.00	0.00	0.00	22
5	0.79	0.71	0.75	48
6	0.53	0.95	0.68	65
7	0.71	0.55	0.62	22
8	0.60	0.83	0.69	52
9	0.40	0.17	0.24	23
10	0.91	0.78	0.84	50
11	0.73	0.71	0.72	38
accuracy			0.67	475
macro avg	0.56	0.58	0.56	475
weighted avg	0.62	0.67	0.63	475

Model 2 Classification Report

	precision	recall	f1-score	support
0	0.33	0.12	0.17	26
1	0.61	0.95	0.74	39
2	0.68	0.52	0.59	29
3	0.86	0.84	0.85	61
4	0.30	0.14	0.19	22
5	0.71	0.31	0.43	48
6	0.63	0.62	0.62	65
7	0.64	0.64	0.64	22
8	0.60	0.73	0.66	52
9	0.78	0.30	0.44	23
10	0.77	0.80	0.78	50
11	0.35	0.76	0.48	38
accuracy			0.61	475
macro avg	0.61	0.56	0.55	475
weighted avg	0.64	0.61	0.60	475

Model Performance Summary- Model 1 vs Model 2

Model 1:

Accuracy: 0.67

Macro Average: Precision = 0.56, Recall = 0.58, F1-Score = 0.56

Weighted Average: Precision = 0.62, Recall = 0.67, F1-Score = 0.63

Model 2:

Accuracy: 0.61

Macro Average: Precision = 0.61, Recall = 0.56, F1-Score = 0.55

Weighted Average: Precision = 0.64, Recall = 0.61, F1-Score = 0.60

KEY INSIGHTS

Accuracy: *Model 1 has a higher accuracy (0.67) compared to Model 2 (0.61).*

Macro Average: *Model 1 outperforms Model 2 slightly in terms of recall (0.58 vs. 0.56), but Model 2 has a higher precision (0.61 vs. 0.56).*

Weighted Average: *Model 1 also performs better in recall and F1-score (0.67 vs. 0.61 for recall, 0.63 vs. 0.60 for F1-score).*

Conclusion

Conclusion- Model Performance

1. **Consistent Top Performers:** Both models effectively identified 'Loose Silky-bent,' 'Common Chickweed,' and 'Small-flowered Cranesbill' as the top three species. This consistency suggests that the features distinguishing these species are well-captured by the models, making them reliable for identifying these specific types of plants.

2. **Challenges with Rare Species:** Both models struggled significantly with 'Black Grass,' 'Common Wheat,' and 'Shepherds Purse.' This could be due to less representative data in the training set for these species, which could lead to poorer generalization on unseen data.

3. **Model 1 vs. Model 2:** While both models showed similar patterns in their top and bottom predictions, subtle differences in the numbers (e.g., slightly better identification of 'Small-flowered Cranesbill' and 'Loose Silky-bent' in Model 2) suggest variances in model sensitivity and possibly feature processing capabilities between the two setups.

Best Practices & Take Aways

- *Data Analysis and Preparation:* In-depth analysis of the dataset to understand distribution patterns and prepare data for modeling is imperative to improve prediction results. This involves image size reduction for computational efficiency and encoding target labels for model training among other processes.

- *Model Development and Evaluation:* Building and refining multiple models to determine the most effective approach for plant species classification includes adjustments to learning rates and the integration of data augmentation techniques to enhance model accuracy and generalizability.

- *Implementation and Impact:* The successful implementation of this model has the potential to significantly impact the agricultural sector by reducing manual labor, increasing yield efficiency, and promoting more sustainable farming practices through precise plant identification.

This project not only represents a significant step forward in agricultural technology but also aligns with broader goals of enhancing sustainability and efficiency in the industry through AI and machine learning innovations.

APPENDIX

Data Background and Contents

Project Overview: Plant Seedling Classification Using Deep Learning

The Aarhus University Signal Processing group, in collaboration with the University of Southern Denmark, has released a dataset containing images of 12 unique plant species. This dataset aims to facilitate the development of a classifier that can accurately determine a plant's species from an image. The data files for this project include:

- images.npy: A collection of 128 x 128 pixel plant images in .npy format, optimized for efficient use.
- Labels.csv: A CSV file containing the species labels corresponding to each image.

The primary goal is to create a Convolutional Neural Network (CNN) to classify plant seedlings into one of the 12 species listed below.

1. Black-grass
2. Charlock
3. Cleavers
4. Common Chickweed
5. Common Wheat
6. Fat Hen
7. Loose Silky-bent
8. Maize
9. Scentless Mayweed
10. Shepherds Purse
11. Small-flowered Cranesbill
12. Sugar beet

Happy Learning !