Revolutionizing Agriculture: A CNN Approach to Seedling Classification

Project :Plant Seedling Classification
Course: AIML
Date: 07/14/2024

Contents

Executive Summary

Enhancing Agricultural Efficiency with Deep Learning

Introduction:

The agricultural industry struggles with the manual identification and classification of plant seedlings, a process that is both time-consuming and labor-intensive. This project addresses this issue by developing a Convolutional Neural Network (CNN) model to classify plant seedlings into 12 categories. The image dataset, provided by the Aarhus University Signal Processing group and the University of Southern Denmark, includes images and labels for 12 distinct plant species. The data was thoroughly loaded, checked for shape, and analyzed to ensure completeness, forming a robust foundation for model development.

Executive Summary Actionable Insights

Address Class Imbalance

- ❑ Implemented oversampling and class weighting to improve accuracy for underrepresented classes like 'Black-grass' and 'Common wheat'

Data Augmentation

- ❑ Continued and expanded data augmentation techniques such as varying degrees of rotation, flips, zooms, and shifts to enhance model robustness and generalization

Regularization Techniques

- ❑ Employed regularization methods such as dropout layers to prevent overfitting and improve generalization to new data

Executive Summary Business Recommendations

Invest in Data Collection

- Gather more data, especially for underrepresented classes, to improve model training and accuracy

Deploy and Monitor

- Deploy the model in a controlled environment and monitor its performance on real-world data to gather feedback and make necessary adjustments

User Training and Education

- Provide comprehensive training and resources to users for effective interpretation and utilization of the model's predictions, enhancing adoption and operational efficiency

Executive Summary

Conclusion :

This project successfully developed a Convolutional Neural Network (CNN) model that classifies plant seedlings with an accuracy of 76.4%, effectively addressing the manual identification challenge in agriculture. Key improvements such as data augmentation and learning rate reduction significantly enhanced model performance. Further steps, including addressing class imbalance, increasing model complexity, and implementing regularization techniques, are recommended to boost accuracy. Strategic investments in data collection, user training, and scalable infrastructure will drive better crop management, reduce manual labor, and increase yields, delivering substantial value to the agricultural industry.

Business Problem Overview and Solution Approach

Problem Statement

As a data scientist, my objective is to tackle the inefficiencies and high labor demands associated with the manual classification of plant seedlings. This project involves developing a Convolutional Neural Network (CNN) model to automate the categorization of plant seedlings into 12 distinct species using image data from Aarhus University Signal Processing group and the University of Southern Denmark. The aim is to employ advanced machine learning techniques to create an accurate, efficient, and scalable solution that reduces the need for manual intervention, enhances classification precision, and facilitates better crop management in the agricultural industry.

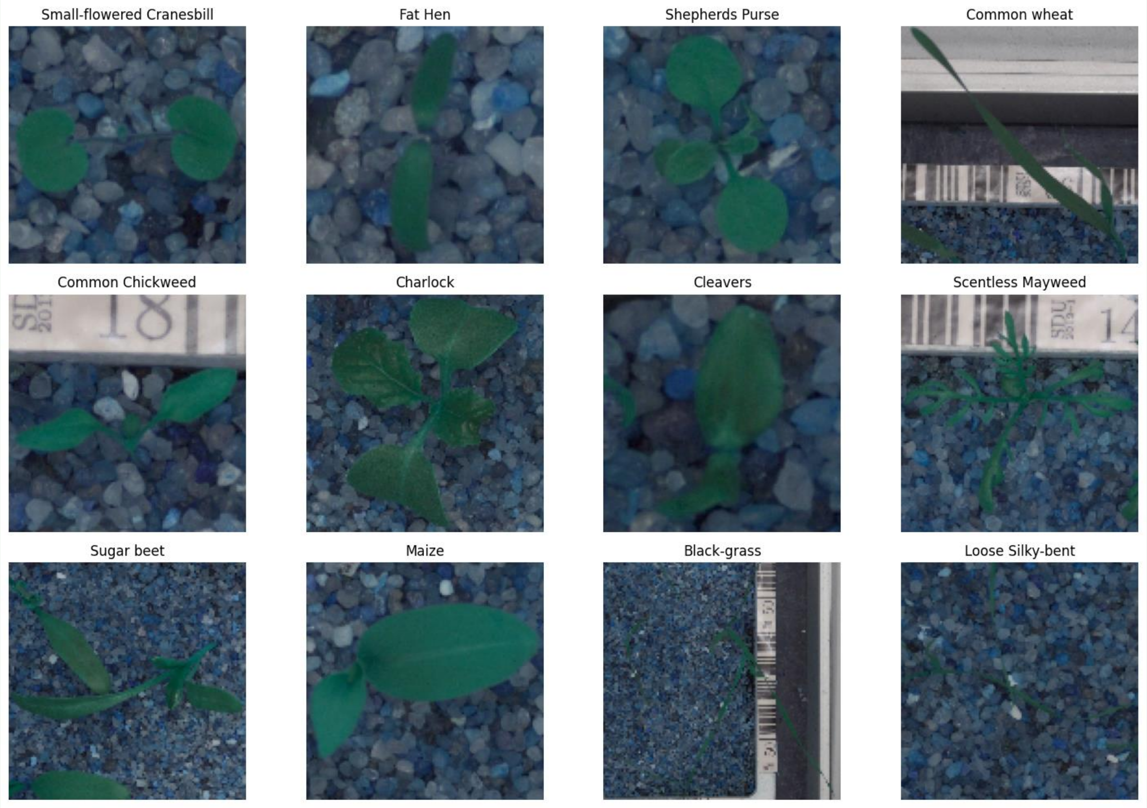
Solution Approach

Data Analysis → Model Development → Implement Solutions → Strategic Execution

Goal

The goal of the project is to create a classifier capable of determining a plant's species from an image

Image Sample of Each Class



Observations:

The images have varying lighting conditions and backgrounds, posing a challenge for the model.

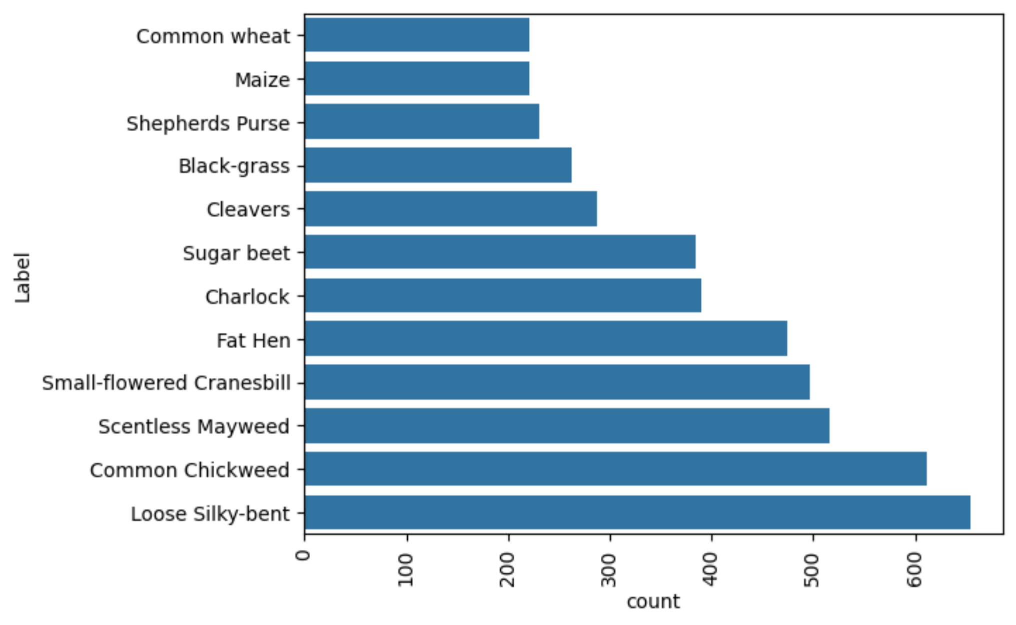
Solution:

1. **Data Augmentation:** Apply techniques such as random brightness adjustment, contrast variation, cropping, rotation, flipping, and zooming to create a more diverse training set

2. **Preprocessing:** Implement preprocessing steps normalization to standardize lighting conditions across images.

3. **Image Filters:** Use filters to reduce noise and enhance important features, improving model robustness against varying backgrounds

Distribution of Plants



Label	Count
Loose Silky-bent	654
Common Chickweed	611
Scentless Mayweed	516
Small-flowered Cranesbill	496
Fat Hen	475
Charlock	390
Sugar beet	385
Cleavers	287
Black-grass	263
Shepherds Purse	231
Common wheat	221
Maize	221

EDA Results and Key Findings

Key Findings

- **Shape:** The dataset consists of 4,750 images (128x128x3) and labels for 12 plant species, each with distinct visual characteristics for CNN differentiation
- **Image Quality:** The images have varying lighting conditions and backgrounds, posing a challenge for the model; data augmentation can help mitigate this issue
- **Class Imbalance:** The dataset is imbalanced, with Loose Silky-bent having the most samples and Maize and Common wheat the least, potentially biasing the model towards more frequent classes
- **Distribution Insights:** Certain species like Loose Silky-bent, Common Chickweed, and Scentless Mayweed are more prevalent, indicates a need for a more effective model training strategy

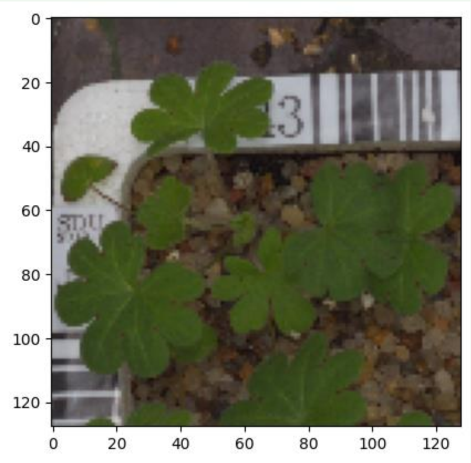
Preparing Data for Model Building

<u>Data Overview:</u>
4,750 images (128x128x3) and labels for 12 plant species

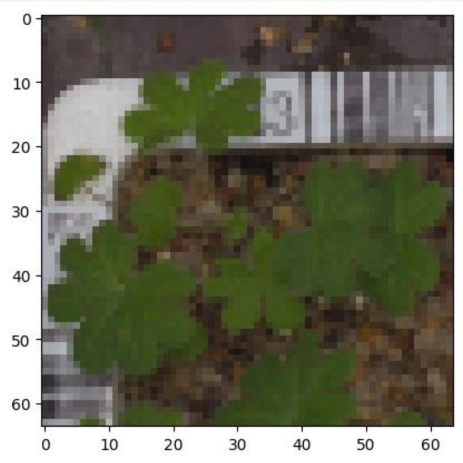
Data Pre-Processing

- **Conversion of Color Scheme**: The images were Converted from BGR to RGB using cvtColor function of OpenCV
- **Resizing Images**: The original images were resized from 128x128 to 64x64 to reduce computational expense during training
- **Data Split**:
  Training Set(3847, 64, 64, 3) (3847,)
  Validation Set (428, 64, 64, 3) (428,)
  Test Set(475, 64, 64, 3) (475,)
- **Encoding Data**: Using LabelBinariser() the target Labels of all three sets were encoded follows (3847, 12), (428, 12), (475, 12)
- **Normalization**: Image pixels were normalized by dividing all the pixel values by 255 to standardize the images to have values between 0-1

Image Resizing



Before (128x128)



After (64x64)

Benefits of Resizing Images

1. **Reduced Computational Load:** Smaller images decrease computational power and memory needs, allowing faster training and the ability to use complex models on standard hardware

2. **Improved Training Speed:** Training on smaller images speeds up learning and enables quicker iterations and experimentation with different models

3. **Feasibility:** Reducing image size to 64x64 makes working with large datasets feasible on limited hardware, avoiding out-of-memory errors

4. **Data Augmentation:** Smaller images facilitate easier application of data augmentation techniques, enhancing model robustness and generalization

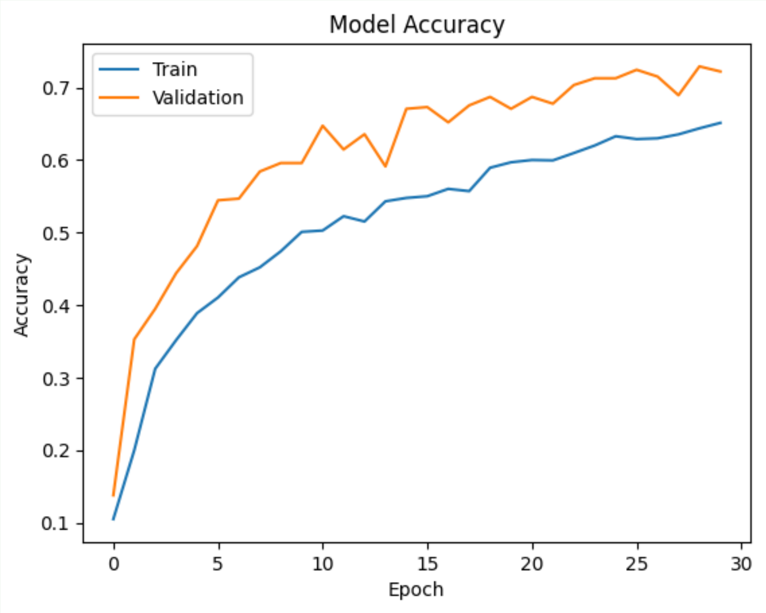
Model 1 (Convolution Neural Network)

<u>Model Summary:</u> Sequential Model fit with 128 filters and kernel size 3x3, padding 'same'

Layer type	Output Shape	Parameter Count
Two-Dimensional Convolution	(None, 64, 64, 128)	3584
Two-Dimensional Max Pooling	(None, 32, 32, 128)	0
Two-Dimensional Convolution	(None, 32, 32, 64	73792
Two-Dimensional Max Pooling	(None, 16, 16, 64)	0
Two-Dimensional Convolution	(None, 16, 16, 32)	18464
Two-Dimensional Max Pooling	(None, 8, 8, 32)	0
Flatten	(None, 2048)	0
Fully Connected or Dense	(None, 16)	32784
Dropout	(None, 16)	0
Fully Connected or Dense	(None, 12)	204

- Total params: 128828 (503.23 KB)
- Trainable params: 128828 (503.23 KB)
- Non-trainable params: 0 (0.00 Byte)

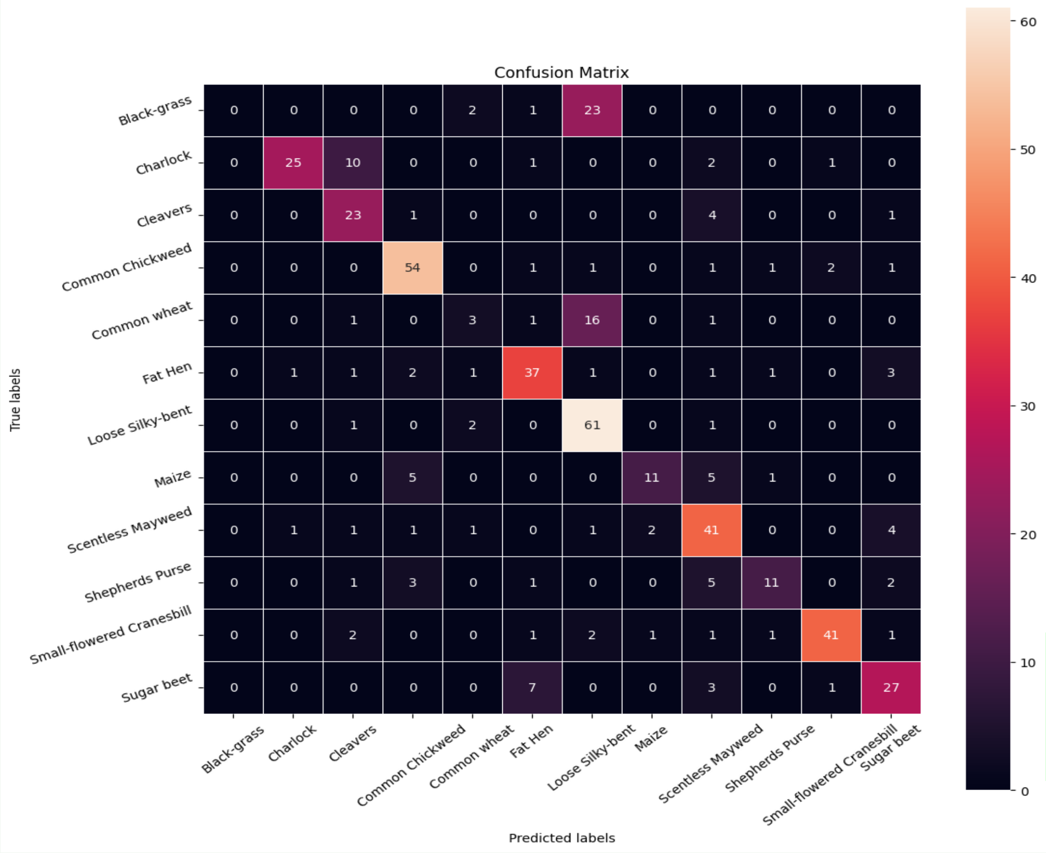
Model Performance on Training Data



<u>Observations from the Model Accuracy Plot</u>

1. **Training vs. Validation Accuracy:** Validation accuracy consistently surpasses training accuracy, suggesting good generalization
2. **Initial Rapid Improvement:** Both accuracies improve quickly in the initial epochs, indicating quick learning of basic patterns
3. **Plateau in Training Accuracy:** Training accuracy plateaus around epoch 20, suggesting the model may have reached its learning capacity
4. **Validation Accuracy Stability:** Validation accuracy shows fluctuations but remains stable, peaking around 0.7, indicating good generalization without overfitting

Model 1 Performance on Test Data



Key Observations

1. High Performance: Classes like Charlock, Common Chickweed, and Loose Silky-bent show strong classification performance

2. Challenges: Classes like Black-grass, Common Wheat, and Maize face significant misclassification issues

3. Misclassification Patterns: Some classes show consistent confusion with specific other classes, indicating a need for improved feature extraction or more training data for those pairs

Recommendations for Further Improvement

Experiment with different learning rates and apply data augmentation and batch normalization

Classification Report of Model 1 on Test Set

Plant Name	Precision	Recall	F1 - Score	Samples
Black-grass	0.00	0.00	0.00	26
Charlock	0.93	0.64	0.76	39
Cleavers	0.57	0.79	0.67	29
Common Chickweed	0.82	0.89	0.85	61
Common wheat	0.33	0.14	0.19	22
Fat Hen	0.74	0.77	0.76	48
Loose Silky-bent	0.58	0.94	0.72	65
Maize	0.79	0.50	0.61	22
Scentless Mayweed	0.63	0.79	0.70	52
Shepherds Purse	0.73	0.48	0.58	23
Small-flowered Cranesbill	0.91	0.82	0.86	50
Sugar beet	0.69	0.71	0.70	38

	Precision	Recall	F1 - Score	Samples
Accuracy			0.70	475
Macro avg	064	0.62	0.62	475
Weighted avg	0.68	0.70	0.68	475

Overall Accuracy:

The model achieved an overall test accuracy of approximately 70.3% indicating that the model is reasonably good at classifying the plant seedlings, but there is room for improvement

Macro and Weighted Averages:

The macro average precision, recall, and F1-score are lower than the weighted averages suggesting that the model performs better on more frequent classes and struggles with less frequent ones.

Performance of Model 1

Overall Accuracy

- The model achieved an overall test accuracy of approximately 70.3% indicating that the model is reasonably good at classifying the plant seedlings

Performance Variability

- The model performs well on some classes with high precision and recall but struggles with others having low precision and recall, likely due to fewer training examples or difficulty distinguishing these classes

Imbalance Impact

- The dataset imbalance may have affected model performance, particularly for underrepresented classes like Black-grass and Common wheat, leading to poor performance on these classes

Confusion Matrix Insights

- The confusion matrix shows that the model often misclassifies Black-grass as Loose Silky-bent and other similar-looking classes.
- There are some misclassifications among classes that may look visually similar, such as Fat Hen and Loose Silky-bent

Macro and Weighted Averages

- The macro average precision, recall, and F1-score are lower than the weighted averages suggesting that the model performs better on more frequent classes and struggles with less frequent ones

Apply Learnings from Model 1

1. **Reduction of Learning Rate:** Improves convergence and avoids plateauing of loss by making finer adjustments during training

2. **Data Augmentation:** Provides more variety to the dataset, enhancing model robustness and generalization

3. **Batch Normalization:** Normalizes inputs of each layer to accelerate training and improve stability

Images after a rotation range of +20 to -20 degrees is applied

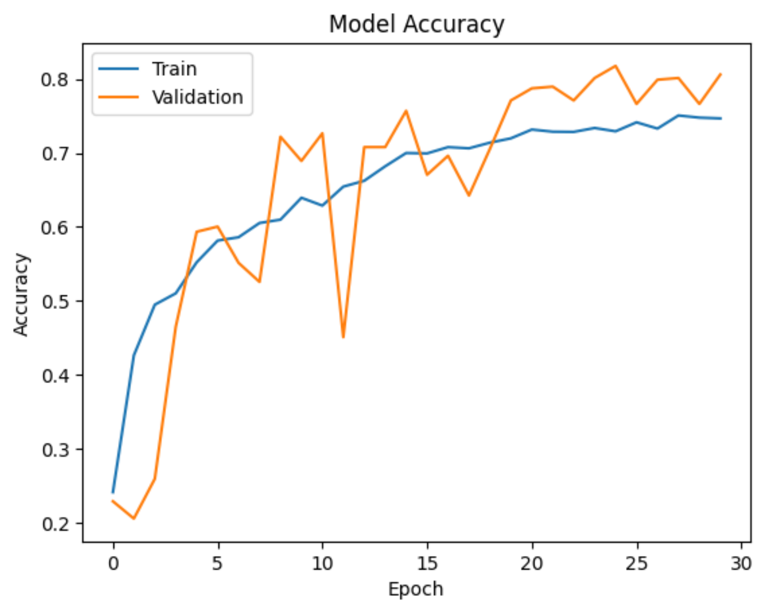
Model 2 (CNN with Improvements)

Model Summary: Sequential Model fit with 64 filters and kernel size 3x3, padding 'same'

Layer type	Output Shape	Parameter Count
Two-Dimensional Convolution	(None, 64, 64, 64)	1792
Two-Dimensional Max Pooling	(None, 32, 32, 64)	0
Two-Dimensional Convolution	(None, 32, 32, 32)	18464
Two-Dimensional Max Pooling	(None, 16, 16, 32)	0
Batch Normalization	(None, 16, 16, 32)	128
Flatten	(None, 8192)	0
Fully Connected or Dense	(None, 16)	131088
Dropout	(None, 16)	0
Fully Connected or Dense	(None, 12)	204

- Total params: 151676 (592.48 KB))
- Trainable params: 151612 (592.23 KB)
- Non-trainable params64 (256.00 Byte)

Model 2 Performance on Training Data



Observations from the Model Accuracy Plot

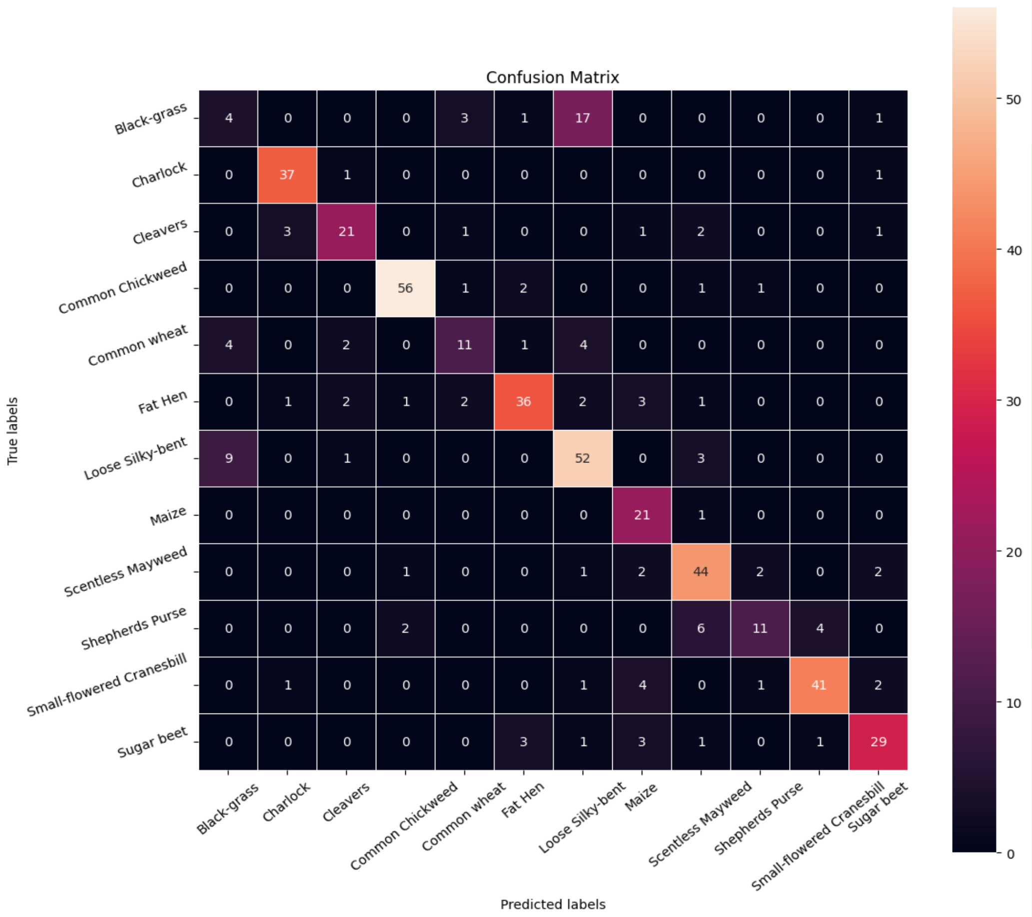
1. **Training vs. Validation Accuracy:** Both training and validation accuracy improve significantly over the epochs, with validation accuracy generally higher than training accuracy, indicating good generalization

2. **Initial Rapid Improvement:** Both accuracies show a rapid increase in the initial epochs, indicating that the model quickly learns the basic patterns in the data

3. **Plateau in Training Accuracy:** After around epoch 20, both training and validation accuracy begin to plateau, indicating that the model is reaching its maximum learning capacity from the given data

4. **Validation Accuracy Stability:** The final accuracy for both training and validation is around 0.8, indicating a relatively good fit with room for further improvement

Model 2 Performance on Test Data



<u>Key Observations</u>

1. **High Performance:** Charlock, Common Chickweed, and Loose Silky-bent show strong classification performance, with most samples correctly classified and only a few misclassifications

2. **Challenges:** Common Wheat and Fat Hen show significant misclassifications, needing better training data or feature extraction. Scentless Mayweed and Small-flowered Cranesbill have some misclassifications, indicating areas for refinement

3. **Misclassification Patterns:** Black-grass and Maize are often misclassified, primarily as other classes, indicating difficulty in accurate differentiation

Model 2 of Classification Report on Test Set

Plant Name	Precision	Recall	F1 - Score	Samples
Black-grass	0.24	0.15	0.19	26
Charlock	0.88	0.95	0.91	39
Cleavers	0.78	0.72	0.75	29
Common Chickweed	0.93	0.92	0.93	61
Common wheat	0.61	0.50	0.55	22
Fat Hen	0.84	0.75	0.79	48
Loose Silky-bent	0.67	0.80	0.73	65
Maize	0.62	0.95	0.75	22
Scentless Mayweed	0.75	0.85	0.79	52
Shepherds Purse	0.73	0.48	0.58	23
Small-flowered Cranesbill	0.89	0.82	0.85	50
Sugar beet	0.81	0.76	0.78	38

	Precision	Recall	F1 - Score	Samples
Accuracy			0.76	475
Macro avg	0.73	0.72	0.72	475
Weighted avg	0.76	0.76	0.76	475

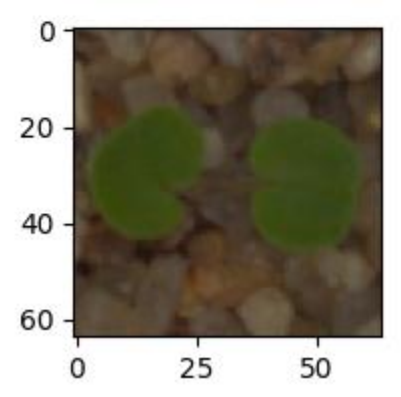
Overall Accuracy:

The model achieved a test accuracy of approximately 76.4%, which is an improvement over the previous model's performance

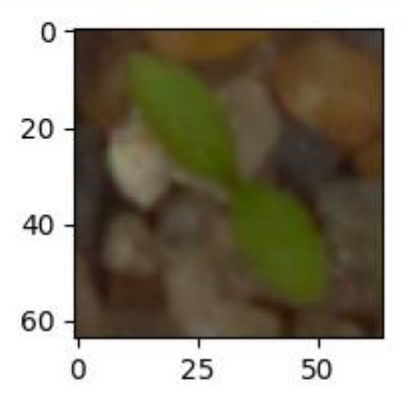
Macro and Weighted Averages:

The macro average precision, recall, and F1-score are slightly lower than the weighted averages, suggesting that the model still performs better on more frequent classes

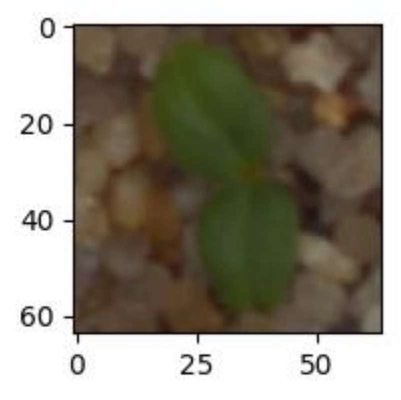
Images of Model 2 Predictions



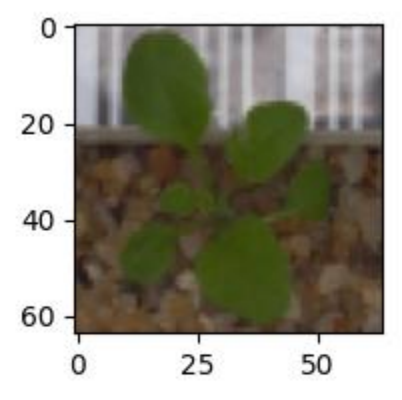
Predicted and true label are both 'Small-flowered Cranesbill



Predicted and true label are both 'Common Chickweed



Predicted and true label are both 'Cleavers'.



Predicted and true label are both 'Shepherds Purse'

The consistency in correctly predicting the labels indicates that the model has learned to identify certain classes accurately and with confidence.

Performance of Model 2

Overall Accuracy

- The model achieved a test accuracy of approximately 76.4%, which is an improvement over the previous model's performance.

Performance Variability

- The model shows improved performance across most classes with high precision and recall, but still has notable misclassifications

Imbalance Impact

- Despite the improvements, class imbalance continues to affect the model's performance, particularly for underrepresented classes like 'Black-grass' and 'Common wheat

Data Augmentation Benefits

- Data augmentation has helped improve generalization, as evidenced by the higher validation and test accuracy

Learning Rate Reduction Impact

- The use of ReduceLROnPlateau has contributed to a more stable learning process, particularly in the later epochs

Model Selection

Chosen Model is Model 2

Model 2 is superior due to its higher accuracy, improved performance across most classes, and better generalization capabilities. It effectively addresses some of the limitations observed in Model 1, making it a more robust choice for the plant seedling classification task

Model 2

Overall Accuracy	Model 2 has a higher test accuracy (76.4%) compared to Model 1 (70.3%)
Class Performance	Model 2 performs better across a wider range of classes, showing improved precision and recall metrics
Generalization	Model 2 demonstrates better generalization, likely due to the implementation of data augmentation and learning rate reduction

Conclusion and Key Takeaways for the Business

Importance of Data Augmentation

1. Data augmentation techniques have been effective in improving model robustness, suggesting further use and refinement could yield additional gains

Scalability and Deployment

2. The model is ready for deployment, but continuous monitoring and periodic retraining with new data are essential for maintaining and improving performance

Business Implications

3. Implementing this model can significantly reduce manual labor, enhance crop management efficiency, and lead to better resource allocation in agriculture

Future Directions

4. Further steps include enhancing data collection, applying advanced augmentation techniques, and exploring more complex model architectures to continue improving performance

Revolutionizing Agriculture: A CNN Approach to Seedling Classification

The End