Plant Seedlings Classification

PGP-AIML-BA-UTA-JUL23-A

Swarup Biswas

Date : 02-02-2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Conclusion
- Appendix

Executive Summary

- **Objective**: The project aimed to leverage Convolutional Neural Networks (CNNs) to classify images of plant seedlings into 12 distinct species. This automation is intended to streamline the agricultural process, reduce manual labor, and enhance crop management efficiency.

- **Data Overview**: The dataset comprised 4750 images across 12 plant species, showcasing substantial variety in terms of shape, size, and color. The data was pre-processed for optimal neural network performance, including converting BGR to RGB, resizing, normalizing, shuffling, and one-hot encoding.

- **Model Development**: Three CNN models were iteratively developed and trained with varying architectures and complexities. Model performance was meticulously tracked, comparing training and validation accuracies and losses to mitigate overfitting and ensure generalization.

Business Insights and Recommendations

Insights and Business Implications:

- Model2 validated with superior accuracy and generalization, suggesting its deployment for real-world agricultural applications.
- The use of CNNs can significantly accelerate the plant classification process, reducing manual workload and allowing agricultural workers to focus on more critical tasks.
- The project's approach could be extended to other domains within agriculture, such as pest detection and soil analysis, further amplifying its potential impact.

Recommendations:

- Model Deployment: Deploy Model2 for real-time plant seedling classification to assist in quick decision-making in agricultural settings.
- Further Data Collection: Continue to gather more diverse data to further enhance the model's accuracy and robustness.
- Continuous Monitoring: Regularly evaluate the model's performance in real-world conditions and retrain with new data to maintain high accuracy.
- Expand Scope: Investigate the application of this model to related agricultural challenges for broader impact.

Business Problem Overview

The agriculture industry requires modern solutions to reduce the extensive manual labor involved in identifying plant seedlings. With the aim of improving crop yields and enhancing higher-level decision-making in agriculture, the project's objective is to leverage artificial intelligence and deep learning to classify plant seedlings efficiently. This technological innovation has the potential to streamline the seedling identification process, leading to increased efficiency, reduced labor, and more sustainable agricultural practices. The challenge is to build a convolutional neural network (CNN) that can accurately categorize images of plants into one of 12 different species, addressing the issue of manual plant sorting and recognition.

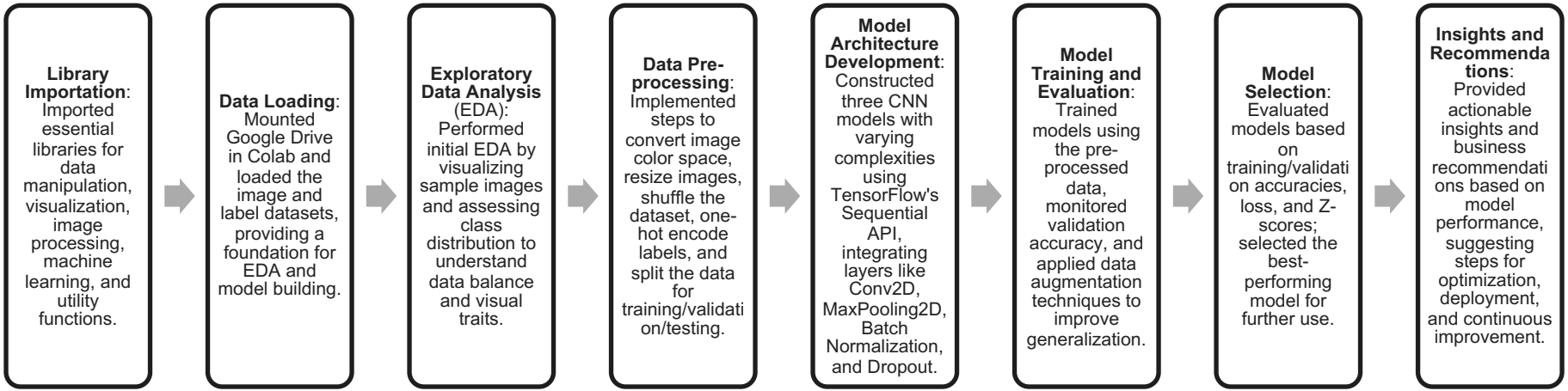
The aim of this project is to Build a Convolutional Neural Network to classify plant seedlings into their respective categories.

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

List of Species

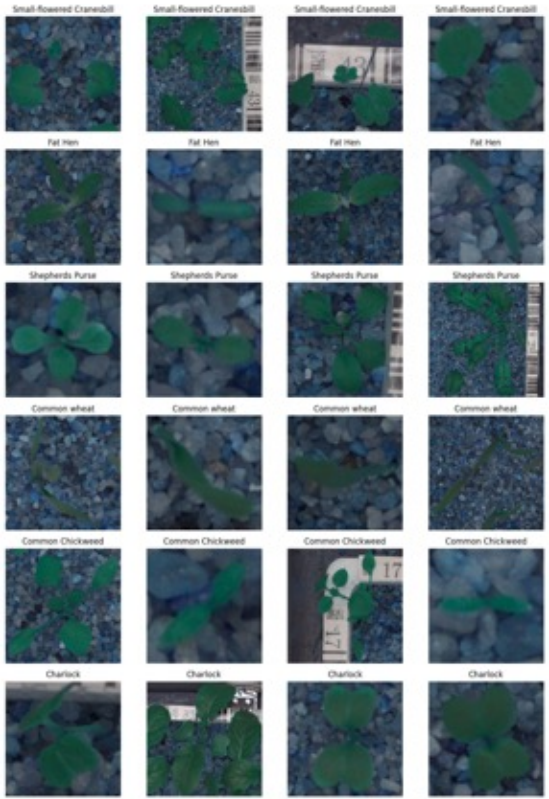
1. Black-grass
2. Charlock
3. Cleavers
4. Common Chickweed
5. Common Wheat
6. Fat Hen
7. Loose Silky-bent
8. Maize
9. Scentless Mayweed
10. Shepherds Purse
11. Small-flowered Cranesbill
12. Sugar beet

Solution Approach



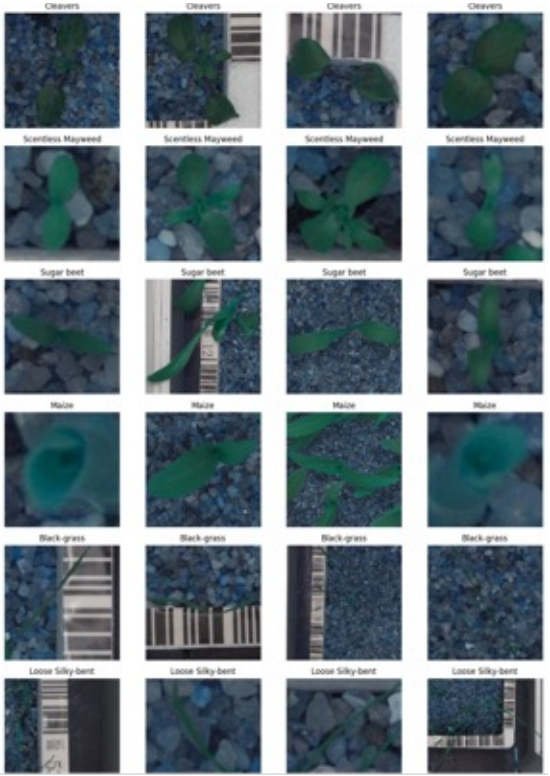
EDA Results

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



The dataset comprises 4750 images, each with a resolution of 128x128 pixels and 3 color channels (RGB), indicating a substantial amount of visual data for model training. Correspondingly, there are 4750 labels, reflecting a one-to-one relationship between images and their categorical labels, essential for a supervised learning task.

Sample Images from Each Category

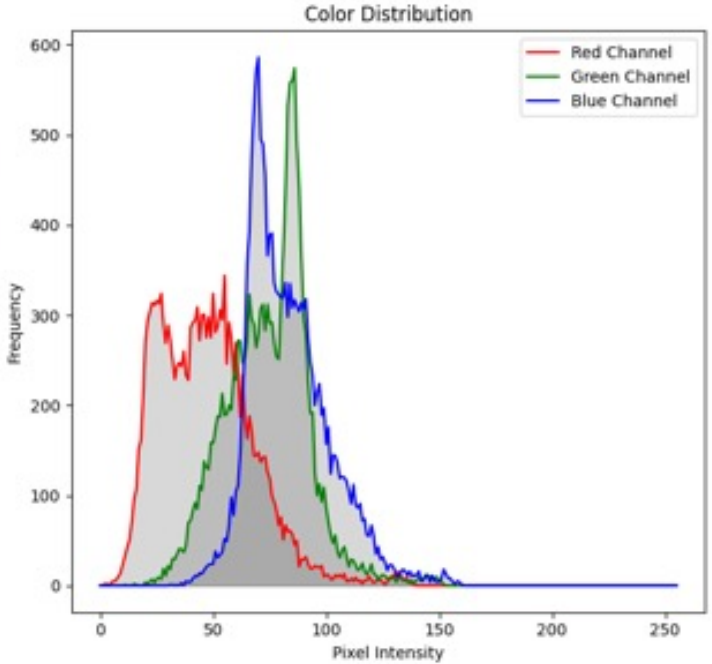


The dataset consists of images belonging to 12 species with the following distribution:

1. Loose Silky-bent: 654 images
2. Common Chickweed: 611 images
3. Scentless Mayweed: 516 images
4. Small-flowered Cranesbill: 496 images
5. Fat Hen: 475 images
6. Charlock: 390 images
7. Sugar beet: 385 images
8. Cleavers: 287 images
9. Black-grass: 263 images
10. Shepherds Purse: 231 images
11. Common wheat: 221 images
12. Maize: 221 images

EDA Results

Random Sample

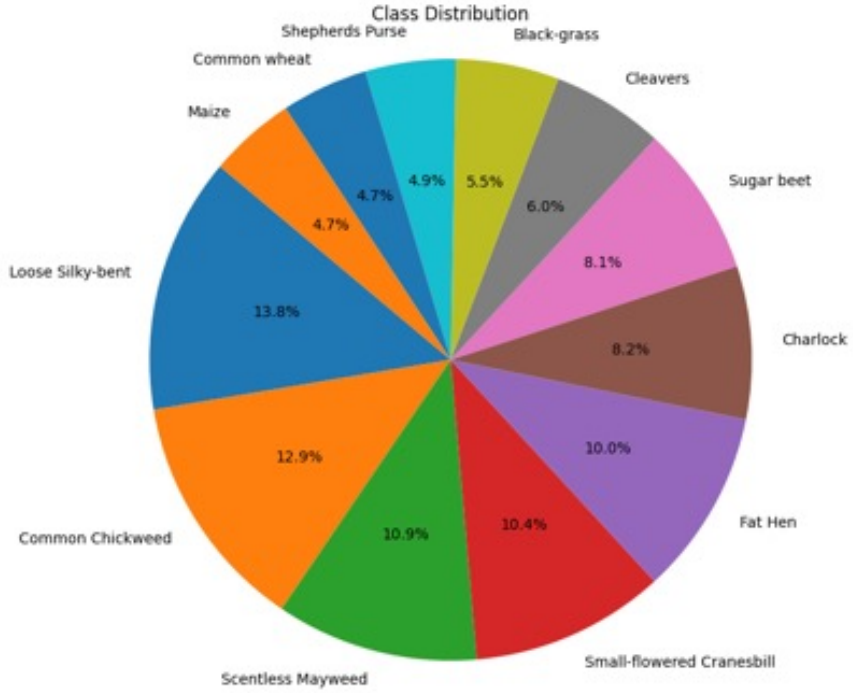


The color distribution plot shows the frequency of pixel intensities for the Red, Green, and Blue channels of a sample image. The green channel peaks at a lower intensity, indicating a dominance of darker green tones, which is consistent with the vegetation present in the image.

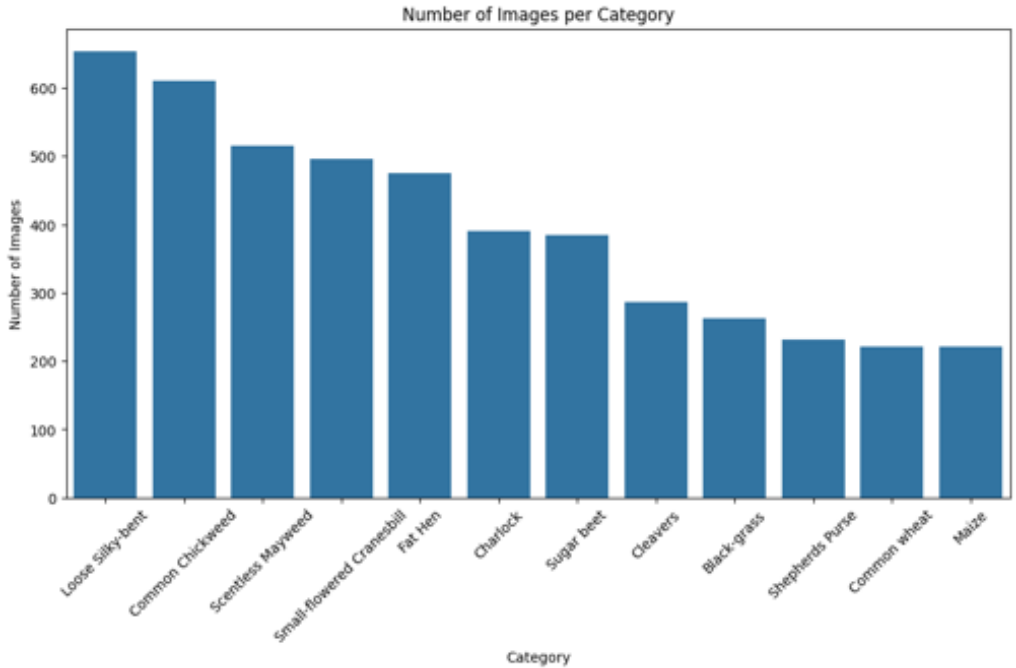
EDA Results

How are these different category plant images different from each other?
Is the dataset provided an imbalance?

The pie chart illustrates the class distribution within the plant seedlings dataset, showing a moderate level of imbalance across the 12 categories. Loose Silky-bent and Common Chickweed are the most represented classes, while Maize and Common wheat have the smallest share, indicating imbalance in data.



The bar chart indicates a moderate imbalance in the dataset, with Loose Silky-bent and Common Chickweed being the most represented classes and Maize and Common wheat being the least.



Data Pre-processing and Preparation Steps

1. **Image Preprocessing and Normalization:**
- The preprocessing function converts images from BGR to RGB color space, ensuring the color channels are correctly ordered for subsequent analysis.
- Images are kept at 128x128 pixels, maintaining a standard size for input into the neural network.
- Pixel values are normalized by dividing by 255.0, scaling them to a [0, 1] range, which aids in the convergence of the model during training.

2. **Dataset Shuffling:**
- The processed images are shuffled using randomly generated indices to ensure the model is exposed to various classes in an unbiased manner during training.

3. **Label Preparation:**
- Labels are one-hot encoded, converting categorical labels into a binary matrix representation required for multi-class classification.

4. **Dataset Splitting with Stratification:**
- The dataset is first split into training plus validation, and test sets, with 10% of the data reserved for testing. Stratification ensures the proportion of classes in each subset is consistent with the original dataset.
- The training plus validation set is further divided into separate training and validation sets, with approximately 11.11% of the original data allocated for validation.
- This stratified split procedure helps maintain an equal class distribution across training, validation, and test sets, which is crucial for training unbiased models.

```
Training set shape: (3800, 128, 128, 3) (3800, 12)
Validation set shape: (475, 128, 128, 3) (475, 12)
Test set shape: (475, 128, 128, 3) (475, 12)
```

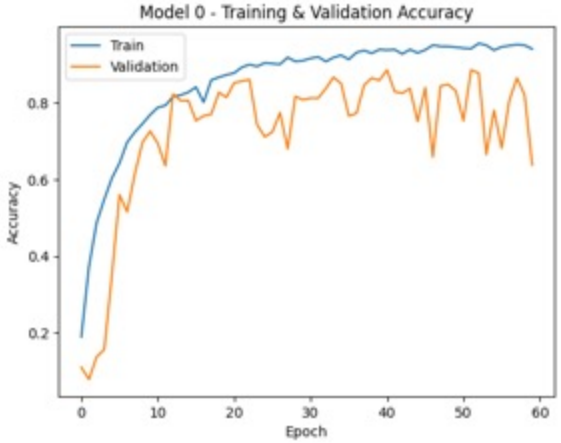
The training set consists of 3800 images and labels, the validation set includes 475 images and labels, and the test set also comprises 475 images and labels, all with the dimensions of 128x128x3 for images and 12 classes for the labels.

Model0 : Base Model Performance Summary

Model0 is a convolutional neural network built using the Sequential API, consisting of four blocks of Conv2D and MaxPooling layers for feature extraction, followed by two Dense layers for classification, with BatchNormalization and Dropout applied to combat overfitting.

Model0 is trained for 60 epochs on the training set without data augmentation, using a batch size of 32 and also validated on a separate validation set.

```
Total params: 230620 (900.86 KB)
Trainable params: 230044 (898.61 KB)
Non-trainable params: 576 (2.25 KB)
```

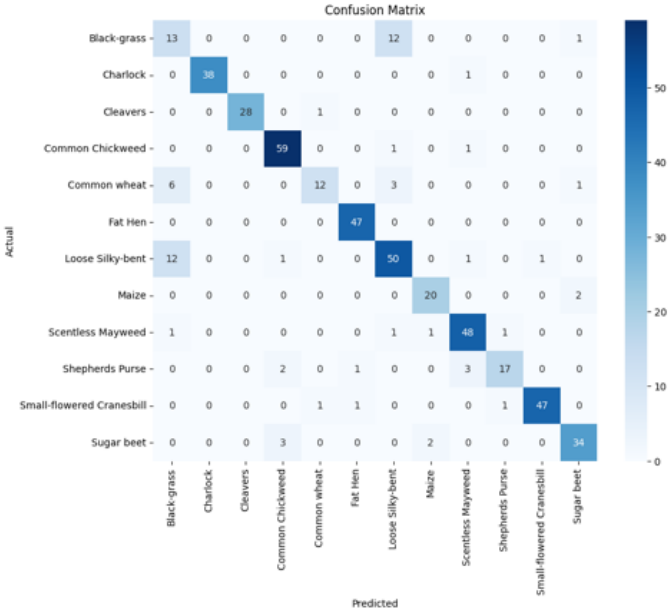


- Model0's training accuracy shows a steady increase and plateaus around 80%, while the validation accuracy fluctuates around 70%, indicating potential overfitting or instability in the model's learning.
- There is a noticeable gap between training and validation accuracy, which suggests that the model could benefit from tuning or more varied training data.
- The confusion matrix for Model0 reveals that the model performs well on certain classes like 'Charlock' and 'Loose Silky-bent' but confuses between classes such as 'Small-flowered Cranesbill' and 'Common Chickweed'.
- The darker diagonal line in the confusion matrix shows correct predictions, but off-diagonal elements indicate misclassifications, with some classes like 'Black-grass' and 'Cleavers' having higher confusion, suggesting room for model improvement.

Model1 : Improvement by Adjusting Learning Rate

Model1 is a Sequential convolutional neural network featuring four convolutional layers with increasing filter complexity and two dense layers, incorporating BatchNormalization and Dropout to enhance performance and mitigate overfitting.
Model1 is trained without data augmentation but with a callback to reduce the learning rate when the validation accuracy plateaus. ReduceLROnPlateau monitors the validation accuracy (monitor='val_accuracy').
If there is no improvement for 3 epochs (patience=3), it reduces the learning rate by a factor of 0.5 (factor=0.5).The learning rate will not be reduced below 0.00001 (min_lr=0.00001).

```
=================================================================
Total params: 230620 (900.86 KB)
Trainable params: 230044 (898.61 KB)
Non-trainable params: 576 (2.25 KB)
_________________________________________________________________
```

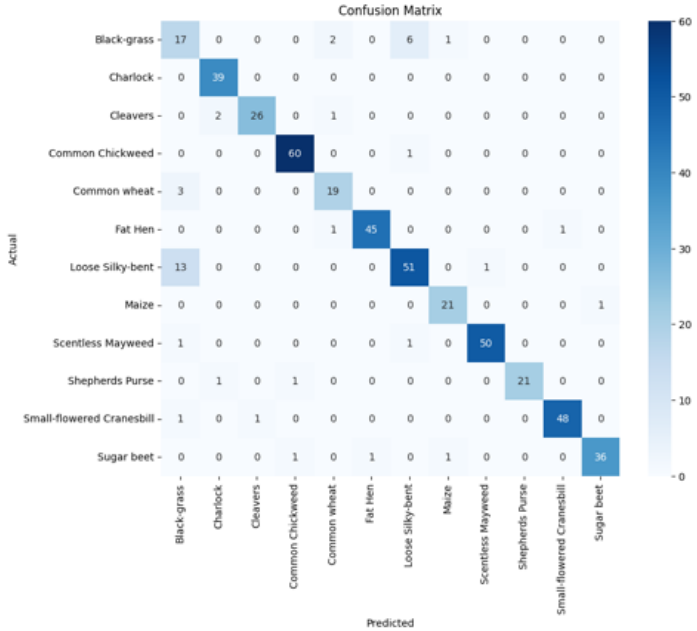


- Model1 achieves a high training accuracy that stabilizes around 90%, and a validation accuracy that closely follows, indicating good generalization with minimal overfitting.
- The accuracy graph demonstrates a consistent trend in training and validation accuracy over 60 epochs, showing the model's reliable learning capability.
- The confusion matrix highlights the model's strength in accurately classifying categories like 'Charlock' and 'Fat Hen' but also reveals some confusion between classes such as 'Black-grass' and 'Loose Silky-bent'.
- The prominent diagonal in the confusion matrix indicates a strong true positive rate for most classes, with some misclassifications suggesting potential areas for model improvement.

Model2 : Improvement by Adjusting Learning Rate & Data Augmentation

Model1 is a Sequential convolutional neural network featuring four convolutional layers with increasing filter complexity and two dense layers, incorporating BatchNormalization and Dropout to enhance performance and mitigate overfitting.
Model1 is trained with Data augmentation with zooming and horizontal flipping is applied to the training data, while the validation data is not augmented, ensuring model robustness and proper evaluation through generators during a 60-epoch training process. And a callback to reduce the learning rate when the validation accuracy plateaus same as Model1.

```
=================================================================
Total params: 230620 (900.86 KB)
Trainable params: 230044 (898.61 KB)
Non-trainable params: 576 (2.25 KB)
_________________________________________________________________
```



- Model2's training and validation accuracy trends closely together throughout the 60 epochs, suggesting a consistent and stable learning pattern with excellent generalization capabilities.
- The training accuracy approaches 90%, and the validation accuracy remains slightly above 80%, indicating the model's robustness despite the complexity of the task.
- The confusion matrix for Model2 displays a strong diagonal, signifying high true positive rates for most classes, with standout performance for 'Charlock' and 'Common Chickweed'.
- Some confusion is observed for the 'Black-grass' and 'Loose Silky-bent' categories, where the model appears to struggle, implying a need for potential model refinement or additional training data for these classes.

Model Performance Summary : Model 2 (Best Model)

	Model	Final Training Accuracy	Final Validation Accuracy	Z-Score Training Accuracy	Z-Score Validation Accuracy	Final Training Loss	Final Validation Loss
0	Model0	0.941053	0.637895	0.026564	-1.407303	0.167808	1.893127
1	Model1	0.967632	0.901053	1.211247	0.582722	0.112044	0.360029
2	Model2	0.912686	0.933036	-1.237811	0.824581	0.285441	0.246768

Analysis

Final Training Accuracy: Model1 has the highest training accuracy (96.76%), indicating it has learned the training data well. However, high training accuracy alone doesn't guarantee a good model if it doesn't generalize well to unseen data.

Final Validation Accuracy: Model2 has the highest validation accuracy (93.30%), indicating it performs best on unseen data among the three models. This metric is crucial as it gives us a good indication of how well the model generalizes.

Z-Score Training and Validation Accuracy: Z-Scores provide a measure of how far off a data point is from the mean in terms of standard deviations. Model1 and Model2 have positive Z-Scores for validation accuracy, with Model2 having the highest, suggesting it's performing above average in terms of validation accuracy compared to the three models.

Final Training and Validation Loss: Model2 also has the lowest validation loss (0.246768), further supporting its ability to generalize well. Lower validation loss is indicative of better performance on unseen data.

Conclusion

Based on the provided metrics, Model2 appears to be the best model. It not only has the highest validation accuracy but also the lowest validation loss, suggesting it has the best generalization capability among the three models. It's also worth noting that Model2's training accuracy is slightly lower than Model1, which might indicate that it's less overfit to the training data compared to Model1, making it a more robust model for new, unseen data.

While Model1 shows excellent training performance, its relatively lower validation accuracy and higher validation loss compared to Model2 suggest it might be overfitting to the training data, making Model2 a preferable choice for generalization. Model0, despite its acceptable training accuracy, falls short significantly in validation accuracy and has the highest validation loss, indicating it's likely overfitting and not generalizing well to new data.

Therefore, in practical applications, Model2 would likely serve as the most reliable model for making predictions on new data, given its balance of good training performance and superior validation performance.

Predicted Label: Small-flowered Cranesbill
True Label: Small-flowered Cranesbill

Predicted Label: Maize
True Label: Maize

Predicted Label: Sugar beet
True Label: Sugar beet

The model has demonstrated precise predictive performance, correctly identifying the species of various plant seedlings such as Small-flowered Cranesbill, Maize, and Sugar beet.

Actionable Insights

1. **Model Optimization**: Continue refining the convolutional neural network models to address misclassifications, particularly between similar-looking plant species like 'Black-grass' and 'Loose Silky-bent'.
2. **Data Augmentation**: Implement further data augmentation strategies to combat overfitting and improve model generalization, particularly for underrepresented classes in the dataset.
3. **Class Balancing**: Address the class imbalance either through data augmentation for minority classes or by using class weights during model training to ensure equal representation and learning.
4. **Transfer Learning**: Explore transfer learning with pre-trained models that have been successful in similar domains to leverage learned features and potentially improve accuracy.
5. **Manual Review Workflow**: Establish a workflow where predictions with low confidence scores are flagged for manual review, thus ensuring high accuracy in critical classifications.
6. **Continuous Model Training**: Set up a system for continuous learning where the model is periodically updated with new data, keeping the classifier current with any changes in plant seedling appearances over time.
7. **Deployment Strategy**: Deploy the model in a user-friendly application to assist agricultural workers in quickly identifying plant seedlings, which will significantly reduce manual labor and improve crop management.
8. **Monitoring and Evaluation**: After deployment, continuously monitor the model's performance and collect feedback for ongoing improvements, ensuring the AI system remains a reliable tool for the agricultural industry.

Great Learning

Happy Learning !