CHAPTER – III

DATA COLLECTION AND PREPROCESSING

3.1 Introduction: An Overview of the Chapter

This chapter delves into the critical processes of data collection and preprocessing, essential for developing the CNN-based system for crop and weed classification. We begin by outlining the methods and sources used to gather a diverse dataset, detailing the conditions under which images were captured and the criteria for selecting representative images of various crop and weed species. In order to increase the diversity of the dataset and strengthen the model's capacity for generalization, the chapter also discusses the annotation process, highlighting the significance of precise labeling. It also outlines data augmentation strategies like rotation, scaling, and color modifications.

Following data collection and augmentation, we discuss the preprocessing steps necessary to prepare the data for model training. This comprises dividing the data into training, validation, and test sets as well as resizing and normalizing it. In order to guarantee a high-quality dataset, we also address issues like imbalanced classes and noisy data and describe techniques like oversampling, undersampling, and data cleaning. This chapter aims to underscore the importance of meticulous data preparation in achieving robust and reliable machine learning outcomes.

3.2 Data Sources

For the development of an effective Convolutional Neural Network (CNN)-based system for crop and weed classification, obtaining a diverse and representative dataset is crucial. This section details the primary and secondary data sources used in this research, highlighting the methods and rationale behind the data collection process.

Primary Data Sources

To ensure the dataset reflects real-world agricultural conditions, primary data was collected through field visits to agricultural regions in West Maharashtra. During these visits, photographs of various crops and weed species were taken under different environmental conditions. In order to gather primary data, high-resolution pictures of the fields were taken, with an emphasis on the various weed species and crop growth stages. The direct field visits allowed for the collection of a wide range of images, providing a robust foundation for the dataset with real-world variability. These images

were taken using a high-quality digital camera to ensure clarity and detail, which are essential for effective training of the CNN model.

Secondary Data Sources

In addition to the primary data collected from field visits, secondary data was sourced using the Google search engine to obtain additional images of the required crops and weed species. This approach helped to expand the dataset, ensuring a comprehensive representation of various plant species. The images retrieved from Google were carefully selected to match the criteria established during primary data collection, such as image quality, relevance, and diversity. By combining these secondary images with the primary data, the dataset was enriched with a broader spectrum of visual scenarios, which is crucial for training a robust and generalized CNN model.

By integrating both primary and secondary data sources, this research ensures a comprehensive and diverse dataset, which is fundamental for developing a reliable and accurate crop and weed classification system. The integration of carefully selected web photographs with field photos from real-world observations yields a well-rounded and large dataset that improves the model's performance in a variety of agricultural scenarios.

3.3 Data Collection Methods

The data collection process for this study involved a systematic approach to gather a diverse and representative dataset of crop and weed images. This section outlines the methods employed for both primary and secondary data collection, detailing the techniques and procedures used to ensure high-quality and relevant data.

Primary Data Collection

The primary data collection involved personal visits to agricultural fields in the West Maharashtra region. The objective was to capture high-resolution images of various crop species and common weeds under natural growing conditions. The following steps were undertaken:

Field Selection: Agricultural fields representing a variety of crops and typical weed species were identified. This selection ensured that the dataset would include a broad range of plant types and growth stages.

Equipment and Setup: A high-quality digital camera was used to take clear and detailed photographs. The camera settings were adjusted to optimize image quality under different lighting conditions. A GPS device was also used to record the locations of the fields for contextual data.

Image Capturing: Photographs were taken at various times of the day to capture different lighting conditions and shadow effects. Multiple angles and distances were used to ensure comprehensive coverage of each plant. Close-up shots were taken to capture fine details, while broader shots provided contextual information about the plant's environment.

Data Logging: Each image was logged with metadata, including the date, time, location, and specific crop or weed type. This information was crucial for organizing the dataset and for future reference during the model training phase.

Secondary Data Collection

To supplement the primary data, secondary images were sourced using the Google search engine. This method was employed to enhance the dataset's diversity and to include images of plants not found in the visited fields. The steps for secondary data collection were as follows:

Keyword Search: Specific keywords related to the required crop and weed species were used to search for images. Keywords included the scientific and common names of the plants, along with terms like "field," "weed," and "crop."

Image Selection Criteria: Images were carefully selected based on quality, relevance, and diversity. High-resolution images with clear visibility of plant features were prioritized. Images depicting various growth stages and environmental conditions were chosen to ensure a comprehensive dataset.

Verification and Annotation: Each selected image was verified for accuracy by cross-referencing with botanical references. Verified images were then annotated, labeling the specific crop or weed species to maintain consistency with the primary data.

Data Integration: The secondary images were integrated with the primary data, ensuring a seamless and organized dataset. Metadata for secondary images included the source URL, date of access, and any additional relevant information.

The work guarantees a strong and high-quality dataset by using these rigorous data gathering techniques, which are necessary for training an efficient CNN model for crop and weed classification. A rich resource for creating a dependable and accurate agricultural monitoring system is provided by the combination of primary field data and carefully chosen secondary photos.

3.4 Description of the Dataset Prepared

The datasets used in this research comprise images of both weed species and crop species, collected from diverse agricultural settings.

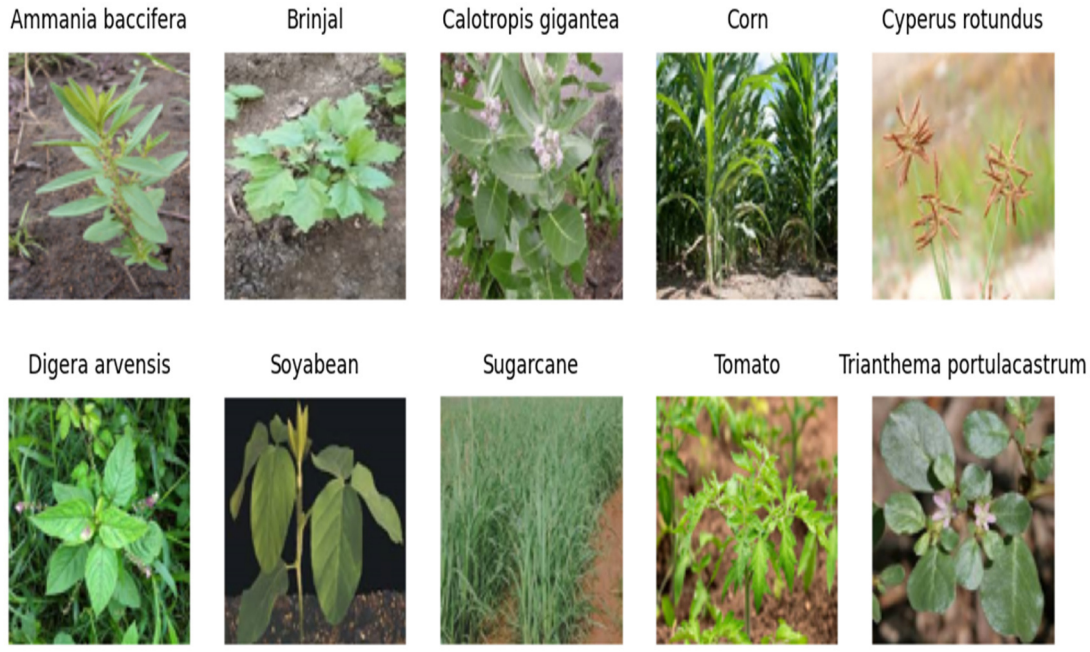

Fig. 3.1 : Random Sample Image of Each Species from the Dataset

3.4.1 Weeds

The weed dataset consists of images representing various common weed species encountered in agricultural fields. The following weed species are included in the dataset:

Cyperus rotundus (Nutgrass)

Known by several names as purple nutsedge or nutgrass, Cyperus rotundus is a perennial plant species that is extensively found in tropical and subtropical areas. It is notorious for its rapid spread and aggressive growth habits, posing a significant challenge to crop cultivation.

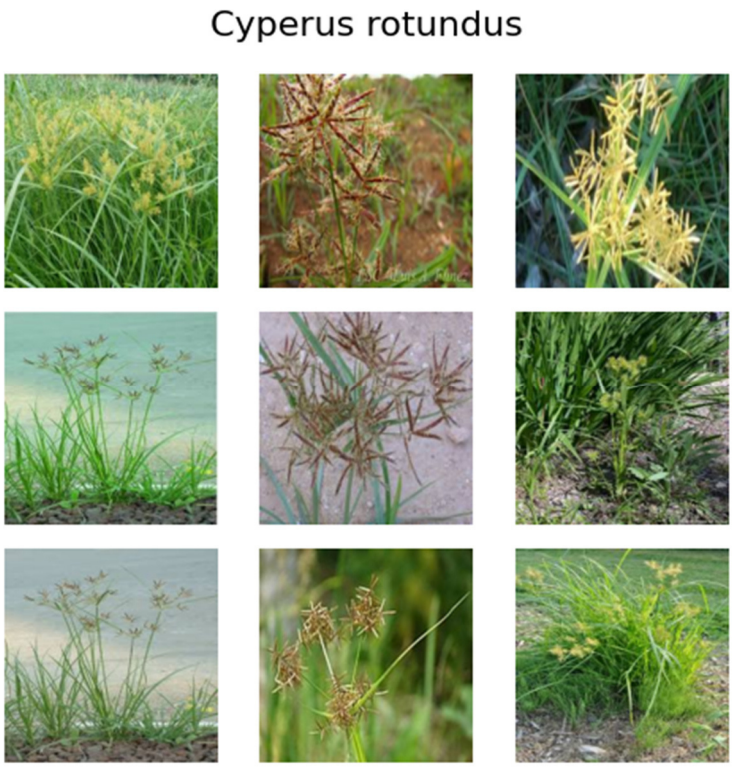

Fig. 3.2 : Random Sample Image of Species Cyperus rotundus from the Dataset

Ammania baccifera (Water willow)

Ammania baccifera, also known as water willow or Bacopa monnieri, is an aquatic weed species commonly found in waterlogged areas such as paddy fields and marshlands. It competes with rice and other crops for nutrients and water, leading to reduced crop yields.

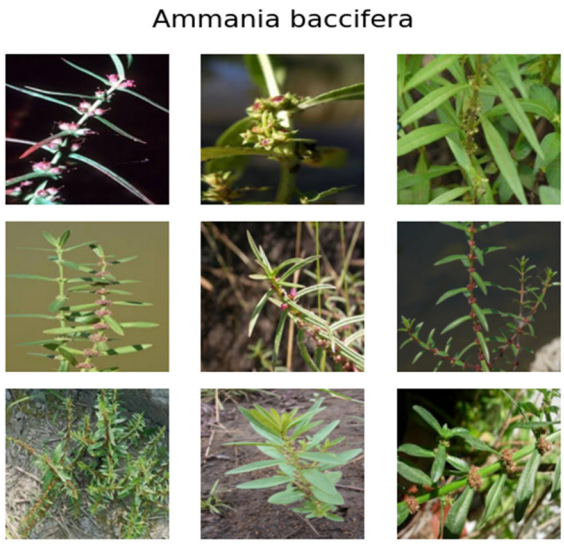

Fig. 3.3 : Random Sample Image of Species Ammania baccifera from the Dataset

Trianthema portulacastrum (Horse purslane)

Trianthema portulacastrum, or horse purslane, is a summer annual weed species prevalent in dry, sandy soils. It thrives in warm climates and is known for its prolific seed production, making it challenging to control in agricultural fields.

Fig. 3.4 : Random Sample Image of Species Trianthema portulacastrum from the Dataset

Digera arvensis (False amaranth)

Digera arvensis, also called false amaranth or red spinach, is a broadleaf weed species found in various agricultural ecosystems. It competes with crops for nutrients and moisture, adversely affecting crop growth and productivity.

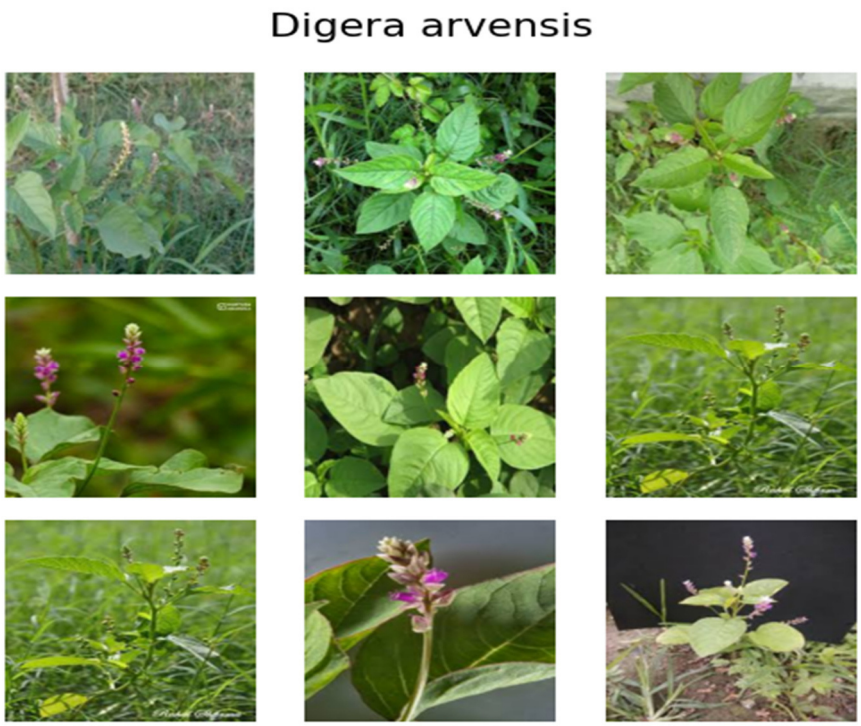

Fig. 3.5 : Random Sample Image of Species Digera arvensis from the Dataset

Calotropis gigantea (Giant milkweed)

Calotropis gigantea is a tropical perennial shrub also referred to as gigantic milkweed or crown flower. It invades agricultural lands and pastures, displacing native vegetation and reducing biodiversity.

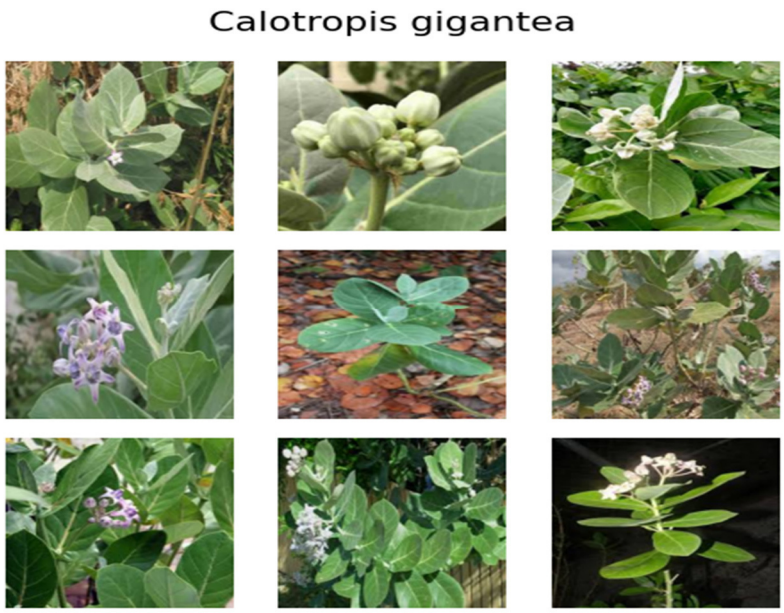

Fig. 3.6 : Random Sample Image of Species Calotropis gigantea from the Dataset

3.4.2 Crops

The crop dataset comprises images representing key crop species cultivated in agricultural fields. These crop species are vital for food security and economic livelihoods in many regions. The following crop species are included in the dataset:

Brinjal (Eggplant)

Brinjal, also known as eggplant or aubergine, is a widely cultivated vegetable crop belonging to the nightshade family Solanaceae. It is grown for its edible fruits, which come in various shapes, sizes, and colors, and are used in diverse culinary dishes.

Corn (Maize)

Zea mays, the scientific name for corn, is one of the most significant cereal crops in the world. In addition to being a basic diet for billions of people, it also provides feed for animals. Corn cultivation is prevalent in diverse agroecological regions, ranging from temperate to tropical climates.

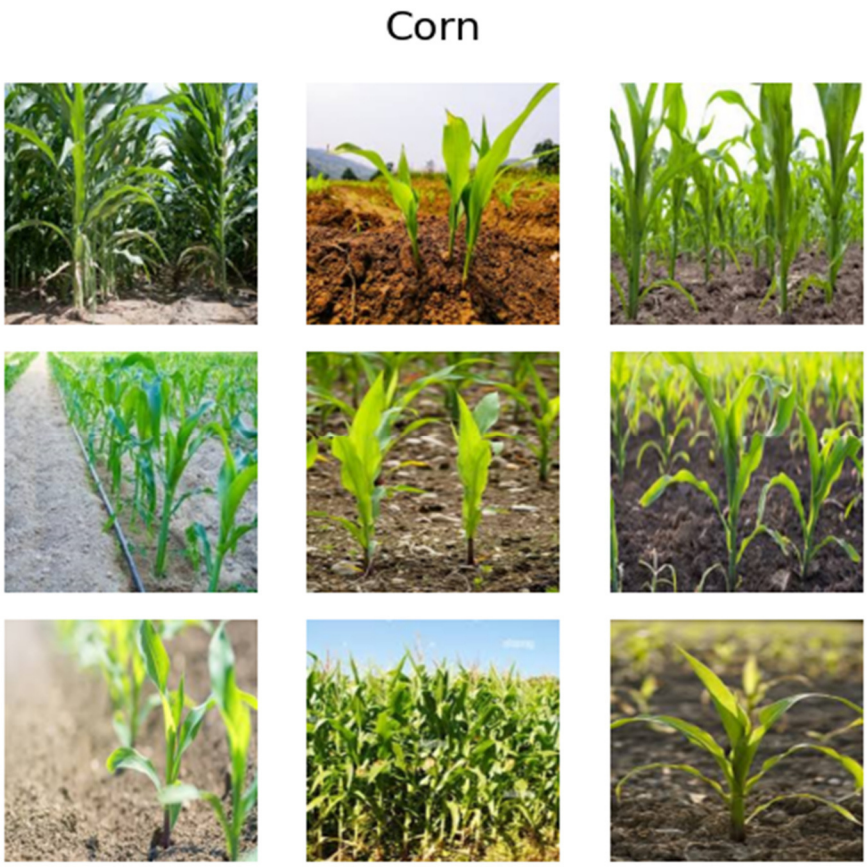

Fig. 3.7 : Random Sample Image of Species Corn from the Dataset

Onion

Onion, botanically known as Allium cepa, is a biennial or perennial vegetable crop cultivated for its edible bulbs. It is a versatile ingredient in various cuisines worldwide and is valued for its pungent flavor and culinary uses.

Soybean

Soybean, or Glycine max, is a leguminous crop species grown for its protein-rich seeds, which serve as a primary source of vegetable oil and protein for human consumption and livestock feed. Soybean cultivation plays a crucial role in global food and feed supply chains.

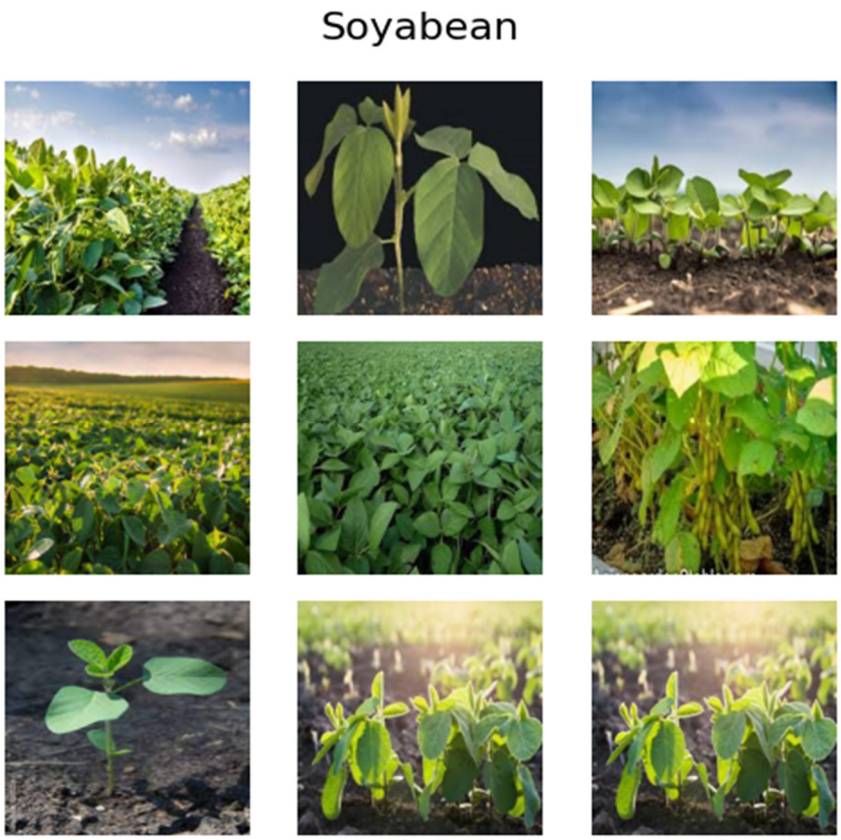

Fig. 3.8 : Random Sample Image of Species Soyabean from the Dataset

Sugarcane

Sugarcane, scientifically known as Saccharum officinarum, is a perennial grass species cultivated for its sweet sap, which is used in sugar production. It is a tropical crop with high water and nutrient requirements, grown primarily for sugar and biofuel production.

The datasets encompass a diverse range of images capturing different growth stages, environmental conditions, and variations in plant morphology for each species. For plant species categorization in precision agriculture applications, these photos form the basis for training and verifying CNN models.

3.5 Distribution of Crop and Weed Images in Dataset

Table 3.1 : Distribution of Crop and Weed Images in Dataset

Sr.No.	Weed Species	Weed Image Counts	Sr.No.	Crop Species	Crop Image Counts
01	Cyperus rotundus	558	06	Brinjal	721
02	Ammania baccifera	570	07	Corn	819
03	Trianthema portulacastrum	487	08	Onion	651
04	Digera arvensis	734	09	Soybean	585
05	Calotropis gigantea	610	10	Sugarcane	595

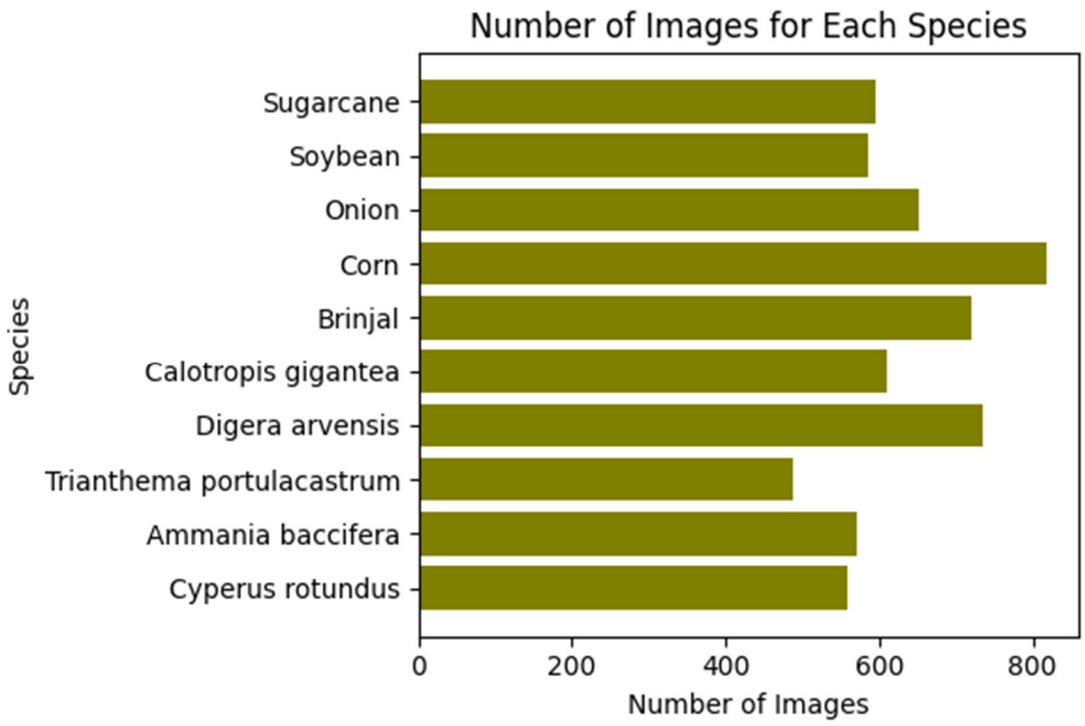

Fig. 3.9 : Bar Chart of Number of Images for Each Species

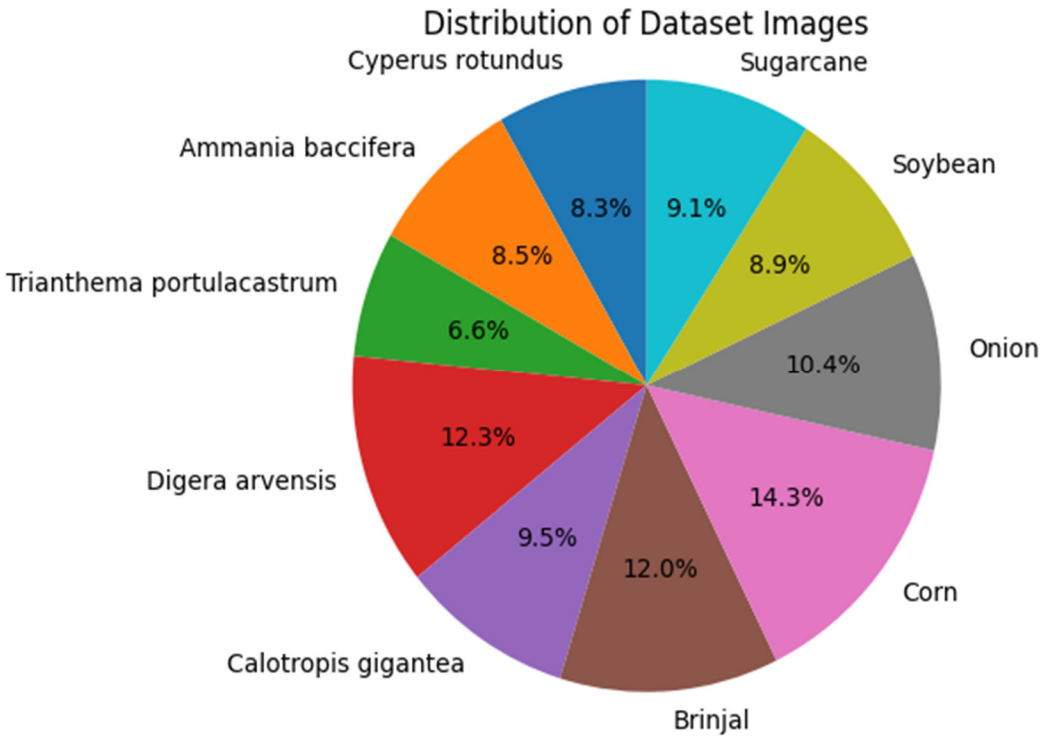

Fig. 3.10 : Distribution of Dataset Images by Percentage

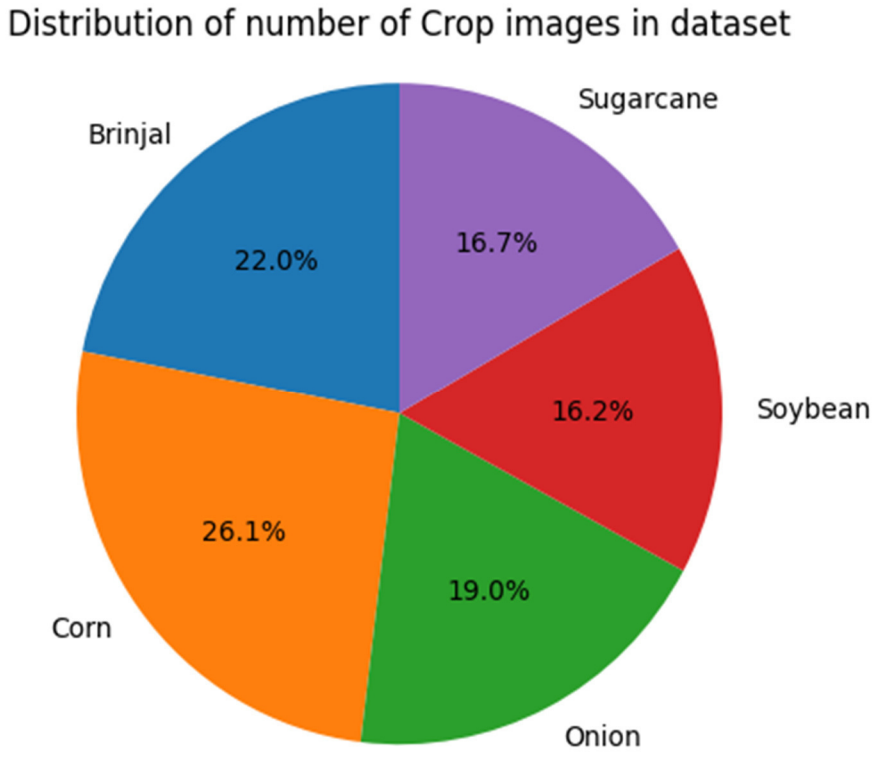

Fig. 3.11 : Distribution of Crop Dataset Images by Percentage

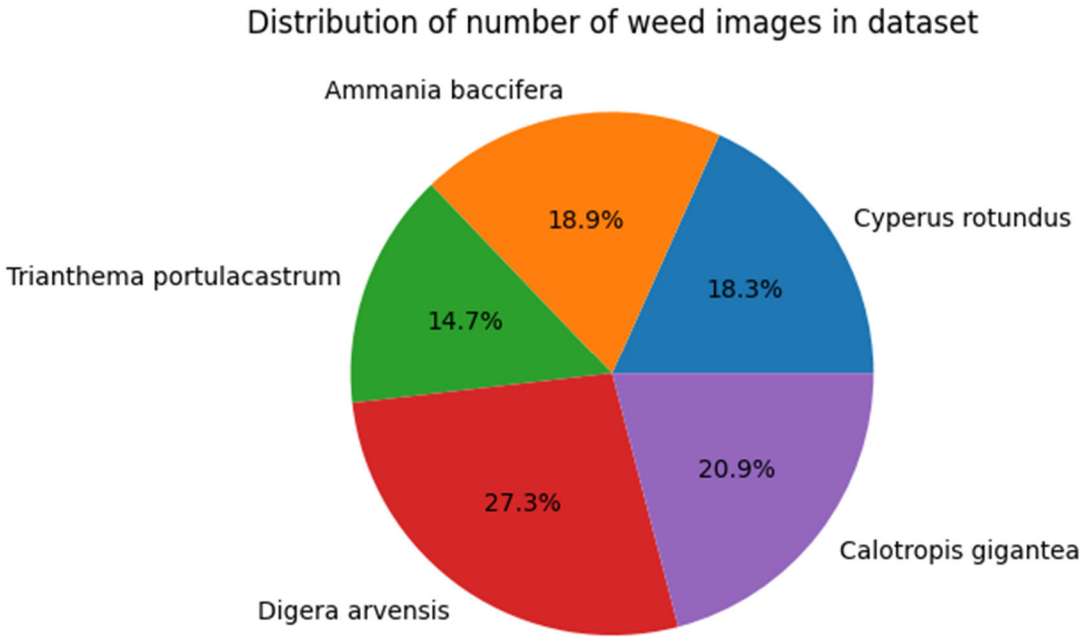

Fig. 3.12 : Distribution of Weed Dataset Images by Percentage

3.6 Data Cleaning

A vital stage in the preparation stage is data cleaning, which guarantees that the dataset is of the highest caliber and devoid of mistakes or inconsistencies that can impair the Convolutional Neural Network (CNN) model's performance. This section outlines the methods and procedures used to clean the data collected from both primary and secondary sources.

1. Removal of Duplicate Images

   The dataset was scanned for duplicate images, which can occur due to multiple captures of the same scene or downloading the same image from different online sources. Duplicate detection was performed using hash-based techniques to ensure each image in the dataset is unique.

2. Correction of Labeling Errors

   Accurate annotation is vital for effective model training. The dataset was reviewed to identify and correct any mislabeled images. This involved cross-checking annotations with botanical references and consulting agricultural experts to verify the correctness of crop and weed labels.

3. Handling Missing Data

   Missing metadata, such as location or time of capture, was filled in where possible. For images where critical information was irretrievably missing, such as species identification, the images were either discarded or flagged for further review.

4. Filtering Low-Quality Images

   Images with poor resolution, blurriness, or excessive noise were removed from the dataset. Quality assessment tools and manual inspection were used to ensure only high-quality images were retained. This step ensures that the CNN model receives clear and informative input data.

5. Balancing the Dataset

   Biased model performance can result from an unbalanced dataset. In order to rectify this, underrepresented types of weeds and crops were found by analyzing the dataset. Techniques such as oversampling of minority classes or data augmentation were employed to achieve a balanced representation of all classes.

6. Normalizing Image Sizes

   All of the photographs were shrunk to a uniform dimension appropriate for the CNN architecture in order to standardize the input data. To prevent image distortion, this preprocessing step involved keeping the aspect ratio intact. Consistent image sizes facilitate efficient processing and model training.

7. Removal of Irrelevant Data

   Images that contained irrelevant content, such as non-agricultural scenes or images with significant portions of background without the target crops or weeds, were removed. This step ensures that the dataset is focused and relevant to the objectives of the study.

8. Data Augmentation for Enhanced Diversity

   Techniques for data augmentation were used to improve the dataset even further. To generate variations of the preexisting photographs, this involved performing operations including rotation, flipping, cropping, and color modifications. Enhancement facilitates the model's capacity to generalize under various circumstances.

9. Verification of Data Integrity

A final review was conducted to ensure data integrity. This included verifying that all images were correctly labeled, free from errors, and consistent with the study's objectives. Automated scripts and manual checks were used to ensure thorough verification.

The work ensures a high-quality dataset that is suitable for training a reliable and accurate CNN model by putting these data-cleaning processes into practice. Effective machine learning relies on clean and trustworthy data, which improves the model's performance and the general dependability of the research findings.

3.7 Data Transformation and Normalization

To guarantee that the dataset satisfies the needs of the Convolutional Neural Network (CNN) model and to enhance the model's training effectiveness and performance, data transformation and normalization are crucial preparation procedures. The steps taken to transform and standardize the gathered data are described in this section.

1. Image Resizing

For each image, the size was adjusted to ensure that it matched the predicted input size of the CNN architecture. Images were reduced to 224x224 pixels for this investigation, a typical size that strikes a compromise between computing efficiency and detail conservation. Resizing helps with batch processing during training and guarantees consistency in the dataset.

2. Aspect Ratio Preservation

While resizing images, the aspect ratio was preserved to avoid distortion. Padding was added where necessary to maintain the original proportions of the images. This is an important step to make sure that the visual properties of the weeds and crops are not changed, as this could have a negative effect on the model's performance.

3. Image Normalization

By dividing each pixel value by 255, the maximum pixel value in an 8-bit image, the pixel values of the image were normalized to a range of 0 to 1. Normalization guarantees that all input features are on a same scale and aids in accelerating the CNN's convergence during training.

4. Color Space Transformation

In cases where the particular model or augmentation approaches demanded it, images were transformed from RGB color space to other color formats, such grayscale or HSV. This transformation can sometimes highlight different features of the crops and weeds, providing additional information for the CNN to learn from.

5. Data Augmentation

Several data augmentation methods were used to increase the dataset's diversity. Among them were:

Rotation: To mimic changes in plant orientation and viewpoint, images were rotated at random angles within a given range. Rotation augmentation improves the model's ability to generalize to plant orientations that are not found in practical settings.

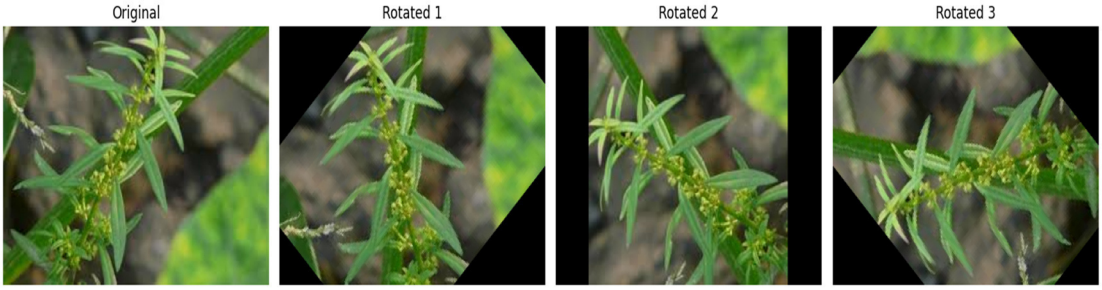

Fig. 3.13 : Sample images by Applying rotation data augmentation to produce rotated images

Flipping: Images were horizontally and vertically flipped with a certain probability to mimic mirror reflections. Flipping augmentation helps improve the model's capacity to identify plants from various angles and positions.

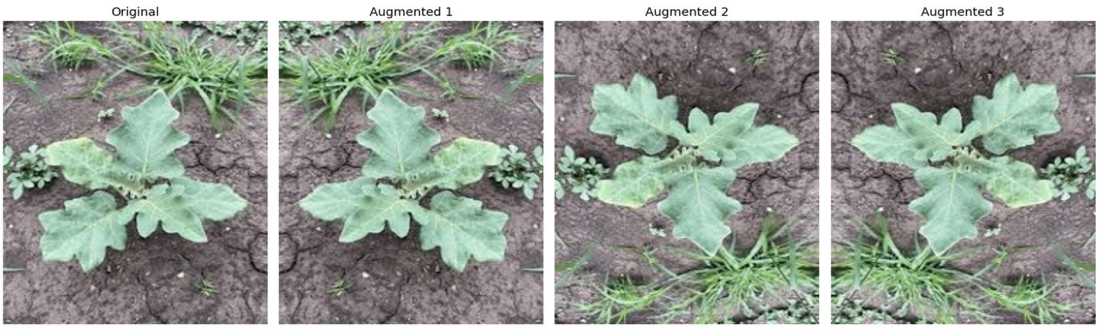

Fig. 3.13 : Sample images by Applying Horizontal and Vertical Flipping data augmentation techniques to produce rotated images

Cropping: Randomly cropping sections of images to focus on different parts of the plants.

Zooming: Random zooming was applied to the images to simulate variations in scale and distance. By learning robust features at various spatial resolutions, zoom augmentation improves the model's capacity to remain scale-invariant and adaptable to changing plant-camera distances.

Fig. 3.15 : Sample images by Applying zooming data augmentation technique to produce rotated images

Brightness and Contrast Adjustment: Random adjustments to brightness and contrast were made to the images to simulate variations in lighting conditions. Brightness and contrast augmentation helps the model learn to distinguish plant features under different illumination levels, making it more robust to lighting variations in real-world environments.

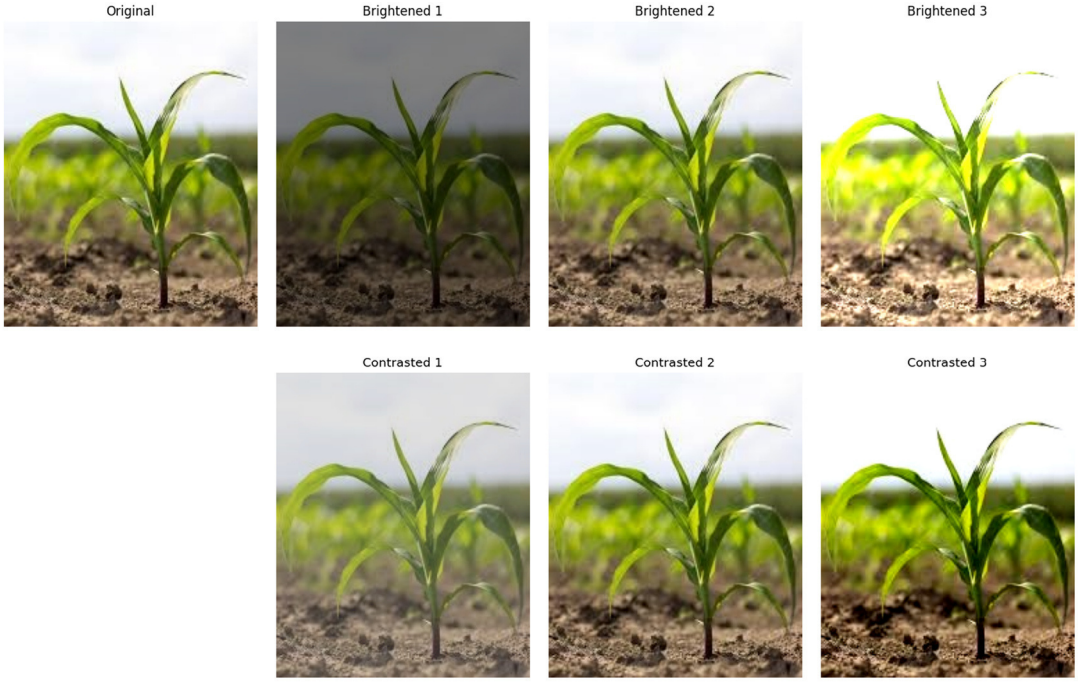

Fig. 3.16 : Sample images by Applying Brightness and Contrast Adjustment data augmentation technique to produce rotated images

6. Standardization

Standardization was also applied by dividing the pixel values for each image by the standard deviation and subtracting the mean. This stage helps to stabilize and expedite the training process by guaranteeing that the dataset has a mean of zero and a standard deviation of one.

7. Data Augmentation Pipelines

Automated data augmentation pipelines were set up using libraries like TensorFlow or PyTorch. Throughout the training phase, these pipelines dynamically apply random changes to the photos, making sure the model is exposed to a large range of image variances.

3.8 Data Splitting Strategy

Three sets of the dataset were created: training, validation, and test. Generally, 10% was utilized for validation, 15% for testing, and 75% of the data was used for training. This division guarantees that the model is trained on most of the available data, and it is tested and validated on data that hasn't been seen yet in order to assess its performance and capacity for generalization.

The study makes sure that the dataset is ready for CNN model training by applying these data transformation and normalization approaches. These preprocessing procedures are essential for improving the model's capacity to learn from the data in an efficient manner, which improves performance in tasks involving the classification of weeds and crops.

The following data-splitting strategy was employed:

75% training, 15% testing, 10% validation.

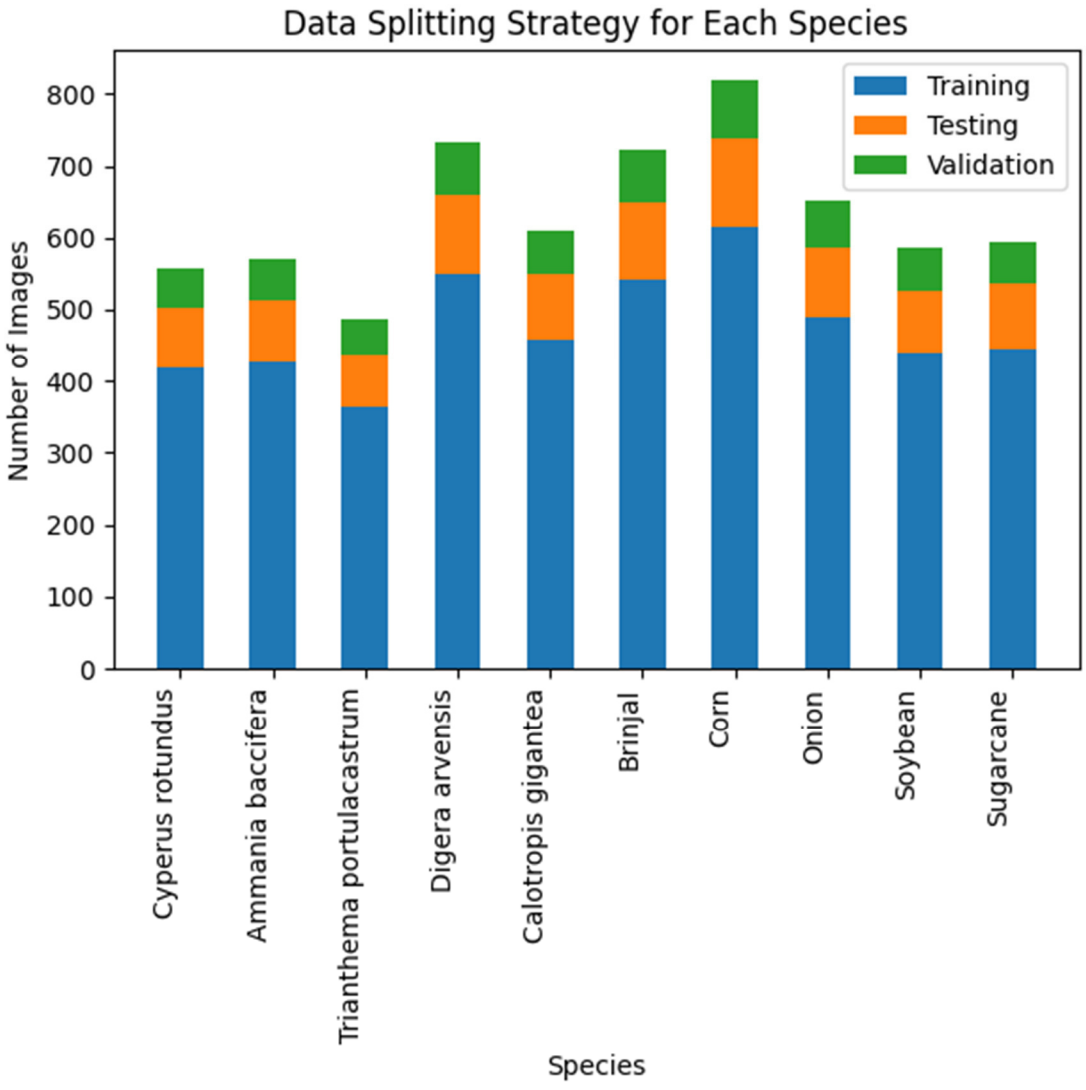

Fig. 3.17 : Data Splitting Strategy for Each Species

Training dataset:

The majority of the dataset is made up of the training set, which is used to train the CNN models. To enable efficient learning and generalization, the training set must

contain a wide variety of images that represent various classes (such as weed species and crop species). To guarantee there is enough data for model training, the training set typically receives 70–80% of the dataset.

Fig. 3.18 : Training Dataset for Each Species

Validation Dataset:

During the training phase, the models' hyperparameters are adjusted using the validation set, which is also used to track how well they perform on untested data. Usually, ten to fifteen percent of the dataset is set aside for the validation set. By offering a separate dataset for evaluating model performance during training, the validation set aids in the prevention of overfitting.

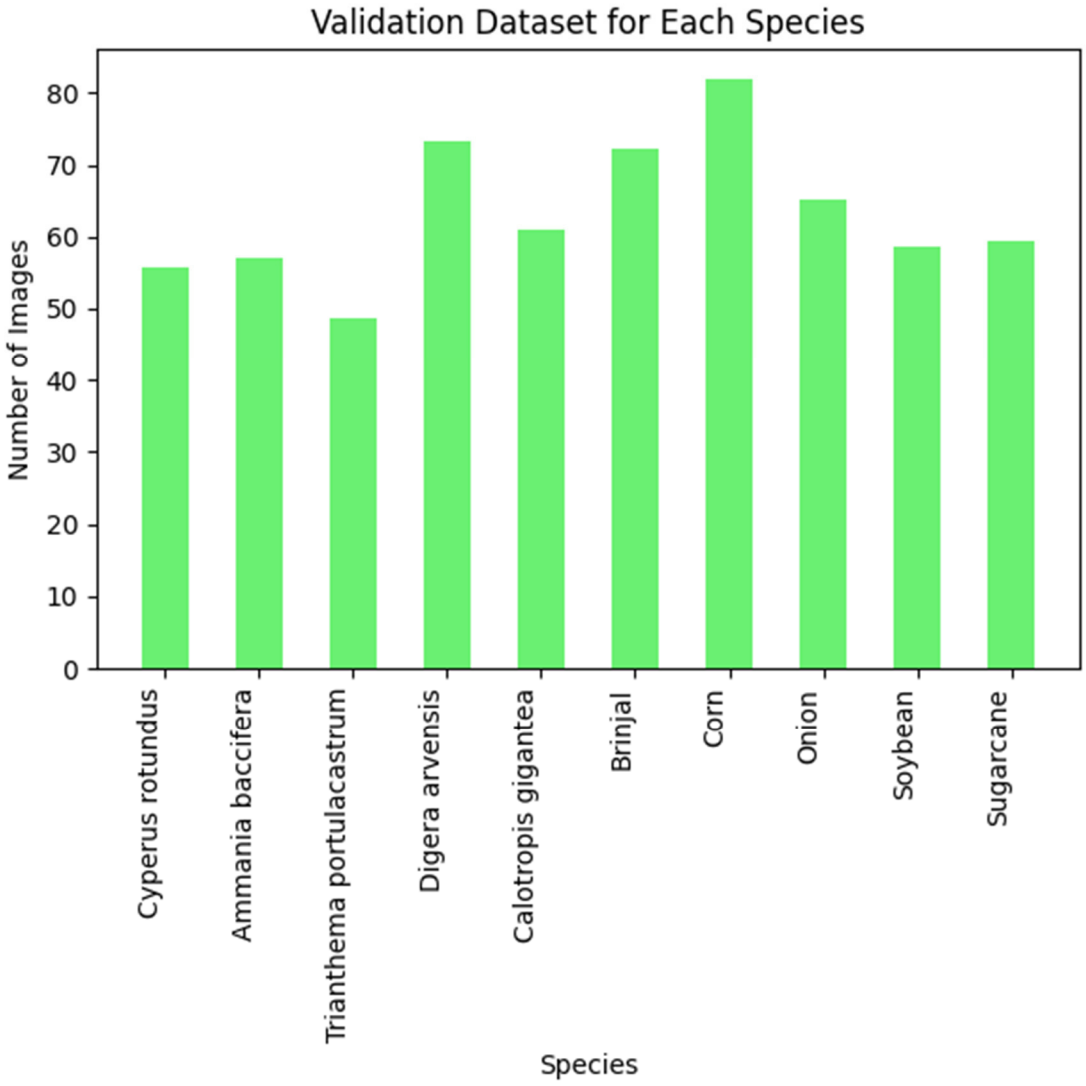

Fig. 3.19 : Validation Dataset for Each Species

Testing Dataset:

The testing set serves as the final evaluation benchmark for the trained models. It consists of completely unseen data that the models have not been exposed to during training or validation. The testing set evaluates the generalization performance of the models on new and unseen samples, providing insights into their real-world applicability. Typically, The testing set receives the remaining percentage of the dataset, which is approximately 10% to 15%.

The procedure of separating the data was carried out while making sure that the training, validation, and testing subsets each preserve an even distribution of images among the various classifications (crop species and weed species, for example).

Before splitting, the dataset was randomly shuffled to eliminate any potential biases in the ordering of the data that can affect how the model is trained and assessed.

By dividing the dataset into discrete subsets for training, validating, and testing, we guarantee a methodical and exacting assessment of the CNN models' efficacy in classifying plant species.

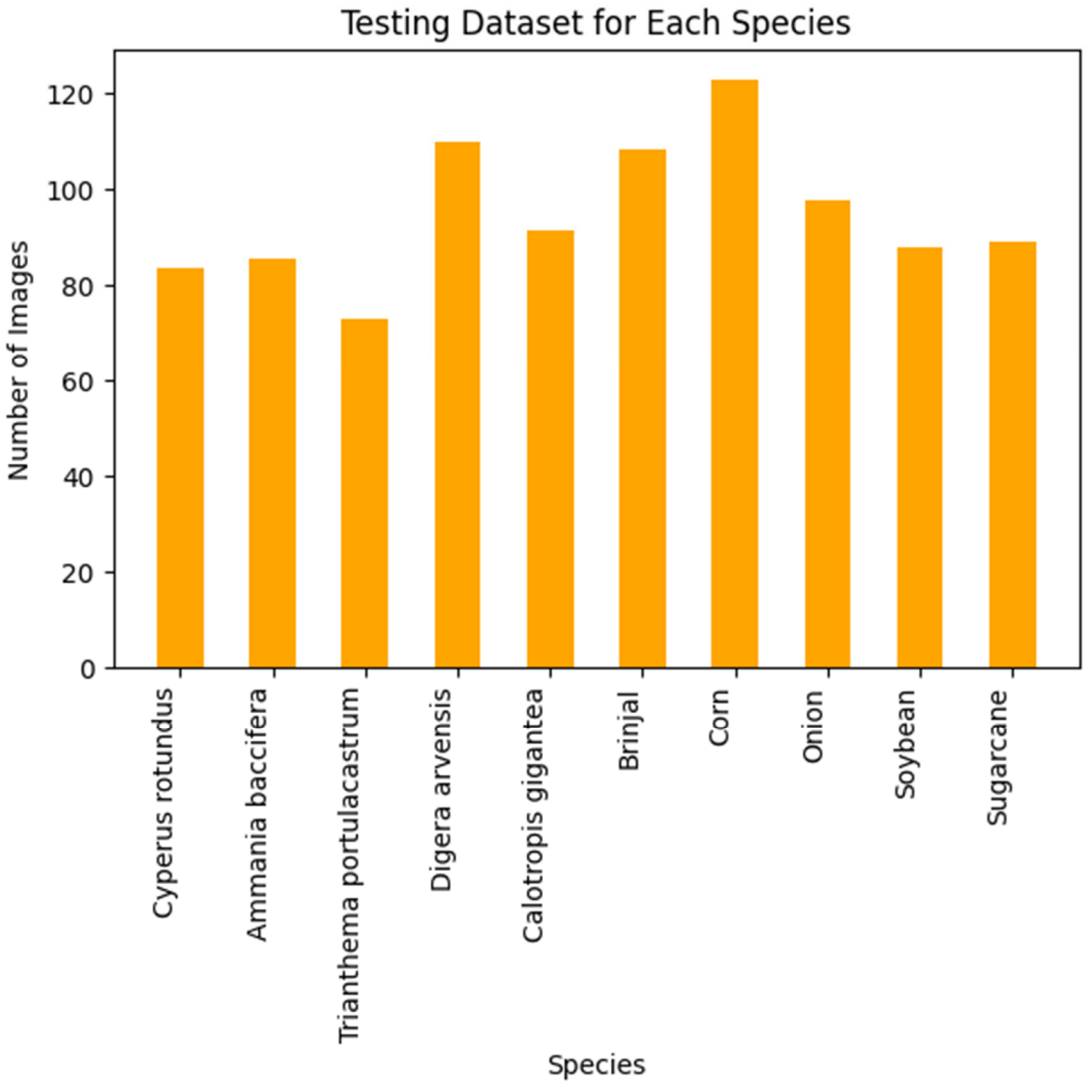

Fig. 3.20 : Testing Dataset for Each Species

75% of the dataset was set aside for training in the data-splitting approach that was used, with the remaining 15% and 10% going to testing and validation, respectively. However, integrating the testing and validation data was required for an efficient model evaluation due to the dataset's relatively modest size. In order to guarantee that the model could be suitably evaluated with a sufficiently large sample size, this

decision was made. The evaluation of the model's performance could be strengthened by combining the testing and validation datasets, reducing the possibility of overfitting or underfitting. Better use of the available data was also made possible by this strategy, which preserved the integrity of the testing and validation processes while optimizing the information utilized for model evaluation. All things considered, merging the testing and validation datasets was a practical way to deal with the limitations caused by the size of the dataset and provide a comprehensive assessment of the model's performance.