Hibiscus Image Dataset: A Lightweight Dataset for Flower Species Classification

1. Venkata Vasanth Sunkara 2. Ayaan Shaikh
3. Shravanpuri Goswami 4. Dhruvi Bharatkumar Patel
Asha M Tarsadia Institute of Computer Science and Technology
guided by Dr. Vishvaji Bakrola

***Abstract*—Accurate classification of plant species is essential for botanical research, agriculture, and conservation efforts. Among flowering plants, Hibiscus is widely cultivated for its medicinal, ornamental, and ecological importance. However, distinguishing between different Hibiscus species can be challenging due to visual similarities, variations in environmental conditions, and the lack of large-scale annotated datasets. In this paper, we introduce the Hibiscus Image Dataset, a publicly available, high-resolution dataset designed for species classification. The dataset consists of 5,100 images across three Hibiscus species: Hibiscus rosa-sinensis, Hibiscus mutabilis, and Hibiscus sabdariffa. Images were collected manually using phone cameras from botanical gardens and nurseries, supplemented with open-source images from online repositories. Each image was preprocessed, resized (252 × 252 pixels), and augmented using transformations such as rotation, flipping, and zooming to improve model robustness.**

We trained a lightweight Convolutional Neural Network (CNN) from scratch to classify the three species. The model was trained using categorical crossentropy loss and optimized with the Adam optimizer (learning rate = 0.0005), achieving high accuracy in species differentiation. The dataset is intended for lightweight deep learning models, making it suitable for mobile applications and real-time plant identification. We provide a detailed analysis of dataset statistics, model performance, and applications. This dataset serves as a benchmark for researchers developing machine learning-based solutions for automated flower classification. Future work includes expanding the dataset by adding more Hibiscus species and increasing the number of images per class, making it suitable for more complex deep learning architectures.

The Hibiscus Image Dataset will be hosted on a dedicated website to facilitate public access and research collaboration. We hope this dataset will contribute to advancements in plant recognition systems and encourage further research in botanical image classification.

I. Introduction

The classification of plant species is a fundamental task in botany, agriculture, and ecological research. Identifying different species accurately is essential for conservation efforts, commercial cultivation, and medicinal applications. Traditionally, plant species classification relies on manual observation by experts, which is both time-consuming and error-prone, especially when species have similar morphological characteristics. With the advancement of computer vision and deep learning, automated plant classification has become a promising alternative, offering fast and accurate species identification. However, one of the major challenges in developing such models is the lack of large, well-annotated datasets for specific plant species.

Hibiscus is a widely cultivated flowering plant with hundreds of species, many of which exhibit high visual similarity, making classification challenging. Despite its widespread presence, there is no publicly available dataset specifically designed for classifying Hibiscus species using machine learning techniques. To address this gap, we introduce the Hibiscus Image Dataset, a high-resolution, manually curated dataset that contains 5,100 images across three distinct Hibiscus species: Hibiscus rosa-sinensis, Hibiscus mutabilis, and Hibiscus sabdariffa. The dataset was collected primarily through manual photography using phone cameras, supplemented with open-source images from online repositories. Each image was resized to 252 × 252 pixels and further augmented through transformations such as rotation, flipping, and zooming to increase diversity and improve model generalization. The dataset is intended to serve as a benchmark for lightweight deep learning models, particularly for mobile-based plant classification applications.

Deep learning models, specifically Convolutional Neural Networks (CNNs), have revolutionized image classification by extracting hierarchical features from images. Unlike traditional machine learning methods, which rely on handcrafted features, CNNs automatically learn spatial patterns, textures, and structures from raw image data. In this study, we designed a lightweight CNN architecture from scratch to classify the three Hibiscus species. The model consists of multiple convolutional layers that extract meaningful features, followed by max-pooling layers to reduce dimensionality. A fully connected dense layer is then used to make predictions, with a softmax activation function determining the final class probabilities. The model is trained using categorical crossentropy loss, optimized with the Adam optimizer (learning rate = 0.0005), and fine-tuned over 20 epochs.

The Hibiscus Image Dataset is a publicly available resource that can be used not only for training machine learning models but also for botanical research, educational purposes, and mobile-based real-time classification systems. This dataset lays the foundation for future research in automated flower classification, and we plan to expand it further by adding more Hibiscus species and increasing the number of images per class to enhance its applicability for larger, more complex deep learning models.

II. Related Work

Several existing datasets focus on plant classification, leaf identification, and general flower recognition. However, most of these datasets do not specifically target Hibiscus species classification. Below, we summarize some relevant datasets:

- Hibiscus Leaf Dataset [1]: A small dataset focusing on Hibiscus leaf classification, primarily for plant disease detection rather than species identification. Available at: https://www.kaggle.com/datasets/sayooj98/hibiscus-leaf-dataset-small.
- Hibiscus Flower Dataset [2]: A dataset containing various Hibiscus flower images, though it is not structured for species classification. Available at: https://universe.roboflow.com/hibiid/hibiscus/dataset/2.
- Plant Species Dataset [3]: A general dataset covering multiple plant species, but it lacks a dedicated focus on Hibiscus flowers. Available at: https://www.kaggle.com/datasets/yousra15b/plant-dataset.
- Flower Image Dataset [4]: A large dataset containing images of different flower species, including Zinnia elegans. However, it does not provide species-level annotation for Hibiscus. Available at: https://www.kaggle.com/datasets/muhammadzulfiga/flower-image-dataset.

While these datasets contribute to plant recognition research, they lack a structured, species-specific dataset for Hibiscus flowers. Our Hibiscus Image Dataset fills this gap by providing a well-annotated, high-resolution collection dedicated to Hibiscus species classification.

III. Dataset Description

The Hibiscus Image Dataset is a curated collection of high-resolution images designed for species classification. It contains a total of 5,100 images, evenly distributed across three Hibiscus species: *Hibiscus rosa-sinensis*, *Hibiscus mutabilis*, and *Hibiscus sabdariffa*. The dataset was primarily collected using mobile phone cameras from botanical gardens and nurseries, supplemented with open-source images to ensure diversity. Each image was resized to 252×252 pixels to maintain uniformity and facilitate efficient training in deep learning models.

A. *Hibiscus rosa-sinensis*

Hibiscus rosa-sinensis is a widely recognized species known for its bright red flowers and large petals. It is commonly used in traditional medicine and as an ornamental plant. The dataset includes 1,700 images of this species, capturing various angles, lighting conditions, and natural variations.

B. *Hibiscus mutabilis*

Hibiscus mutabilis, commonly known as the Confederate Rose, is known for its unique ability to change color throughout the day. It typically starts as white in the morning, turns pink in the afternoon, and deepens to red by evening. The dataset includes 1,700 images of this species to capture its natural variations.

Fig. 1: Sample images of Hibiscus rosa-sinensis

Fig. 2: Sample images of Hibiscus mutabilis

C. *Hibiscus sabdariffa*

Hibiscus sabdariffa, also known as Roselle, is widely cultivated for its edible calyces, which are used in herbal teas and food products. This species has distinct dark red or maroon-colored petals, making it visually different from the other two species in the dataset. A total of 1,700 images of *Hibiscus sabdariffa* are included in the dataset.

The dataset is balanced, with an equal number of images per class, ensuring fair training and evaluation, making it ideal for developing lightweight machine learning models for accurate Hibiscus species classification.

Fig. 3: Sample images of Hibiscus sabdariffa

IV. Methodology

A. Image Collection

The images in the Hibiscus Image Dataset were primarily captured using mobile phone cameras in natural lighting conditions from botanical gardens, home gardens, and nurseries. Due to the limited availability of all three Hibiscus species in a single location, efforts were made to source images from multiple regions. Images were taken at varying angles, distances, and under different lighting conditions to introduce diversity. Since acquiring a sufficient number of images manually was challenging, additional images were obtained from open-source online repositories. Care was taken to ensure that the dataset only contained high-quality images that clearly depict the flower structure without excessive noise or background obstructions.

B. Data Preprocessing and Augmentation

To standardize the dataset and improve model generalization, all images were resized to a fixed resolution of 252 × 252 pixels. Additionally, various data augmentation techniques were applied to artificially expand the dataset and introduce variations in orientation, scale, and lighting. The following augmentations were performed:

- Rotation at multiple angles to account for different viewing perspectives.
- Horizontal and vertical flipping to increase dataset diversity.
- Width and height shifting to simulate positional variations.
- Zooming and shear transformations to enhance feature robustness.

The dataset was then split into three subsets: 70% for training, 15% for validation, and 15% for testing.

C. Model Architecture and Training

To evaluate the dataset, a Convolutional Neural Network (CNN) model was developed from scratch. The architecture consists of:

- Three convolutional layers with ReLU activation functions for feature extraction.
- MaxPooling layers to reduce spatial dimensions and retain important features.
- Fully connected dense layers for classification.
- Softmax activation in the final layer for multi-class classification.

Figure 4 illustrates the model structure.

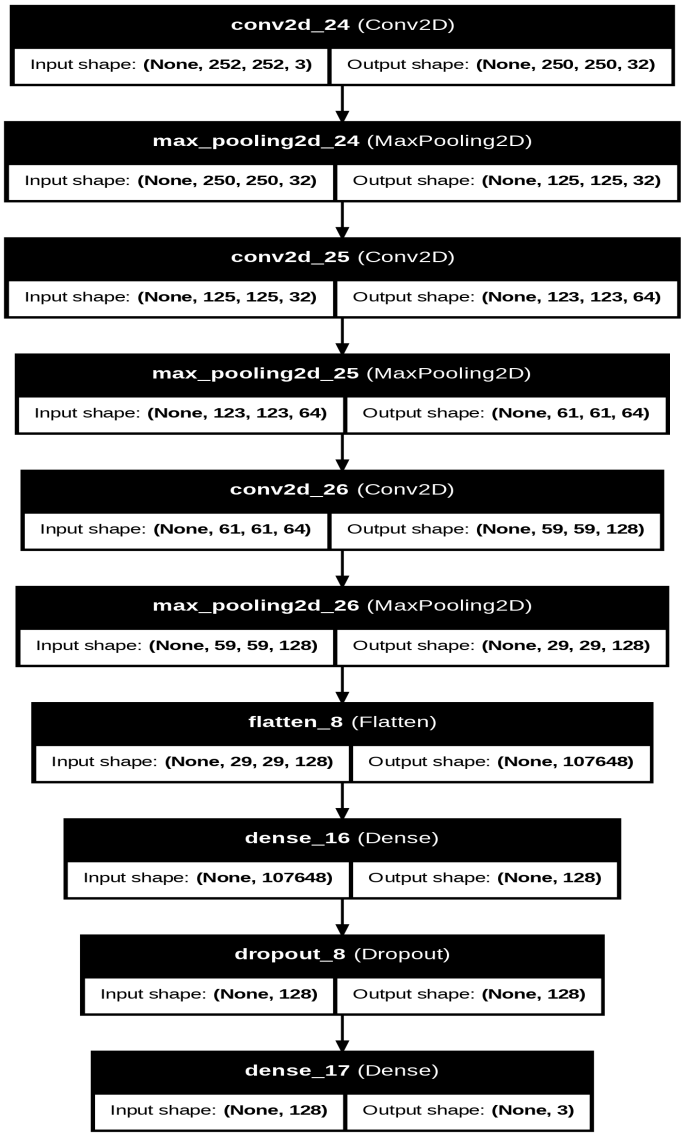

Fig. 4: CNN Model Architecture for Hibiscus Species Classification

The model was compiled using the Adam optimizer with a learning rate of 0.0005. The loss function used was categorical crossentropy, as the dataset contains three classes. The model

was trained for 20 epochs with a batch size of 32. To prevent overfitting, early stopping was implemented, monitoring validation loss with patience set to 5 epochs.

D. Evaluation and Testing

The trained model was evaluated on the test dataset, which consisted of 15% of the total images. Performance metrics such as accuracy and loss were tracked for both the training and validation sets. The final model was assessed by plotting accuracy and loss curves to observe learning patterns over epochs. The testing phase determined how well the model generalized to unseen data, ensuring reliable classification of Hibiscus species.

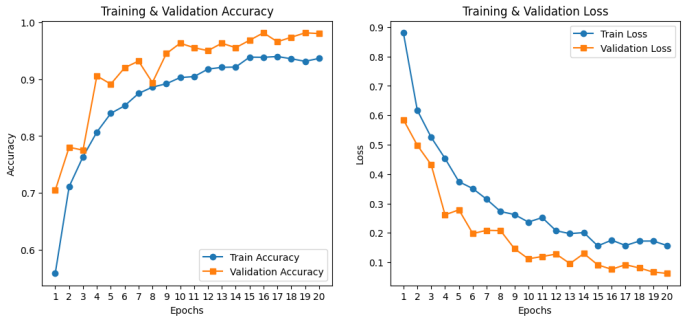

Fig. 5: Training and Validation Accuracy/Loss over Epochs

V. Dataset Statistics

The Hibiscus Image Dataset consists of a total of 5,100 images, evenly distributed across three Hibiscus species. The dataset was carefully balanced to ensure that each class contains an equal number of samples, reducing the risk of model bias toward any particular species. Table I provides an overview of the dataset distribution.

TABLE I: Dataset Statistics

Species	Number of Images	Percentage
Hibiscus rosa-sinensis	1,700	33.3%
Hibiscus mutabilis	1,700	33.3%
Hibiscus sabdariffa	1,700	33.3%
Total	5,100	100%

Since each species has an equal number of images, the dataset is well-suited for training classification models without the risk of class imbalance. This ensures that the model does not develop a preference for any particular class, leading to fair and accurate classification results.

The dataset is divided into three subsets:

- 70% (3,570 images) for training
- 15% (765 images) for validation
- 15% (765 images) for testing

This split ensures that the model is trained effectively while maintaining a separate validation set for hyperparameter tuning and a dedicated test set for performance evaluation. The dataset distribution is optimized for machine learning applications, particularly convolutional neural networks (CNNs), which require diverse and well-balanced training samples for robust classification.

VI. Applications

The Hibiscus Image Dataset is a valuable resource for multiple applications in machine learning, botanical research, and real-world implementations. Some of the key applications include:

- **Machine Learning and Deep Learning Research:** The dataset enables the training and evaluation of convolutional neural networks (CNNs) and other classification models for species recognition.
- **Mobile and Web-Based Plant Identification Systems:** The dataset can be integrated into mobile applications that allow users to identify Hibiscus species in real-time using smartphone cameras.
- **Botanical Studies and Horticulture:** Researchers and horticulturists can use the dataset to analyze morphological differences between species and assist in plant taxonomy studies.
- **Educational Purposes:** The dataset serves as an educational tool for teaching machine learning concepts and plant classification techniques.
- **Pretraining for Other Flower Recognition Models:** Due to its balanced structure, the dataset can be used as a pretraining resource for broader flower classification models.

By making the dataset publicly available, we aim to facilitate further research and development in the fields of computer vision, botany, and mobile-based plant recognition systems.

VII. Conclusion

This paper introduced the Hibiscus Image Dataset, a well-structured and publicly available dataset for Hibiscus species classification. It consists of 5,100 high-resolution images evenly distributed across three species: *Hibiscus rosa-sinensis*, *Hibiscus mutabilis*, and *Hibiscus sabdariffa*. Images were collected from various sources, ensuring diversity and quality, and were preprocessed and augmented to enhance model generalization.

A Convolutional Neural Network (CNN) was trained on the dataset, demonstrating its effectiveness for species classification. The balanced dataset structure ensures unbiased learning, making it valuable for plant recognition, botanical research, and mobile-based identification systems.

Future improvements will focus on expanding the dataset with more images and species while incorporating additional annotations. The dataset will be hosted on a dedicated website for public access, supporting further research in computer vision and plant classification.

Acknowledgment

The authors would like to thank all contributors who assisted in data collection, annotation, and dataset validation.

REFERENCES

[1] Hibiscus Leaf Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/sayooj98/hibiscus-leaf-dataset-small.

[2] Hibiscus Flower Dataset, Roboflow Universe. [Online]. Available: https://universe.roboflow.com/hibiid/hibiscus/dataset/2.

[3] Plant Species Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/yousra15b/plant-dataset.

[4] Flower Image Dataset, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/muhammadzulfiga/flower-image-dataset.