Development of an AI-powered herbarium image pre-identification system in Violaceae: focusing on Korean species

Hyeonji Moon [P1] , Su-Jeong Han [2], Jaesung Lee[2], and Sangtae Kim [1]

[1]Dept. of Biology, Sungshin Women's University, Seoul 01133, Rep. of Korea;
[2]Dept. of Artificial Intelligence, Chung-Ang University, Seoul 06974, Rep. of Korea

amborella@sungshin.ac.kr

IBC 2024 XX International Botanical Congress Madrid Spain

Introduction

The era of global meta-herbarium (Davis, 2022): a paradigm shift in botanical study

- Plant specimens serve as essential repositories of morphological, biogeographical, and ecological data and play a key role in basic botanical research such as plant taxonomy, phylogenetics, and ecology, as well as in practical applications such as oriental medicine and pharmaceutical developments.
- Driven by technological advances, the world's leading herbaria are rapidly digitizing their collections, making specimens accessible in digital form and more readily available for research and related applications worldwide (Table 1).

Table 1. Representatives of digital herbaria

Country	Digital Herbarium	# of images[a]	Web site
USA	iDigBio (Integrated Digitized Biocollections)	33,040,765	https://www.idigbio.org/
Australia	AVH (The Australasian Virtual Herbarium)	7,201,432	https://avh.chah.org.au/
France	MNHN (Muséum national d'Histoire naturelle)	5,568,392	https://science.mnhn.fr/institution/mnhn/collection/p/item/search/form
Brazil	Reflora Virtual Herbarium	4,135,048	https://reflora.jbrj.gov.br/reflora/herbarioVirtual/ConsultaPublicoHVUC/ResultadoDaConsultaNovaConsulta.do?lingua=en
India	IVH (The Indian Virtual Herbarium)	136,658	https://ivh.bsi.gov.in/
China	CVH (The Chinese Virtual Herbaium)	66,113	https://www.cvh.ac.cn/index.php
Russia	MW (Moscow University Herbarium)	37,553	https://moscow.depo.msu.ru/

Misidentified specimens in herbaria

- Data quality, expressed as the accuracy of identification, is a key factor in the success of subsequent studies.
- However, large herbaria still hold many misidentified specimens, and re-identifying them requires much time and effort from taxon-specific experts.

Application of deep learning for preliminary identification and evaluation of preidentified herbarium sheets

- Transfer learning initializes a network by pre-training it on a broad set of available data and then fine-tuning it with sparser, domain-specific data (Carranza-Rojas et al., 2017).
- Previous deep-learning studies to identify plant specimens have typically used pre-trained models such as ResNet (de Rutio et al., 2022; Hussein et al., 2022; Shirai et al., 2022).
- The PlantCLEF 2022 competition (https://www.imageclef.org/PlantCLEF2022) is an example of a recent effort to identify species from large-scale specimen images, where four million plant specimen images were classified into 80,000 classes (Goëau et al., 2022).

Violaceae: the first target taxon

- This research aims to develop a fast and efficient automatic re-identification system for herbarium specimens of Korean plants using curated datasets and convolutional neural networks, which will ultimately be used to improve the data quality of herbarium specimens dramatically.
- ***Viola*** (Violaceae) are distributed in ca. 660 taxa worldwide (ca. 40 in Korea). Their specimens in Korea have high misidentification rates due to 1) morphological similarity among species and 2) seasonal or environmental variation within species.
- As a preliminary study before the full-scale project, we tried automatic identification using transfer learning on a ResNet-based model for 36 species of Korean violets (Violaceae).

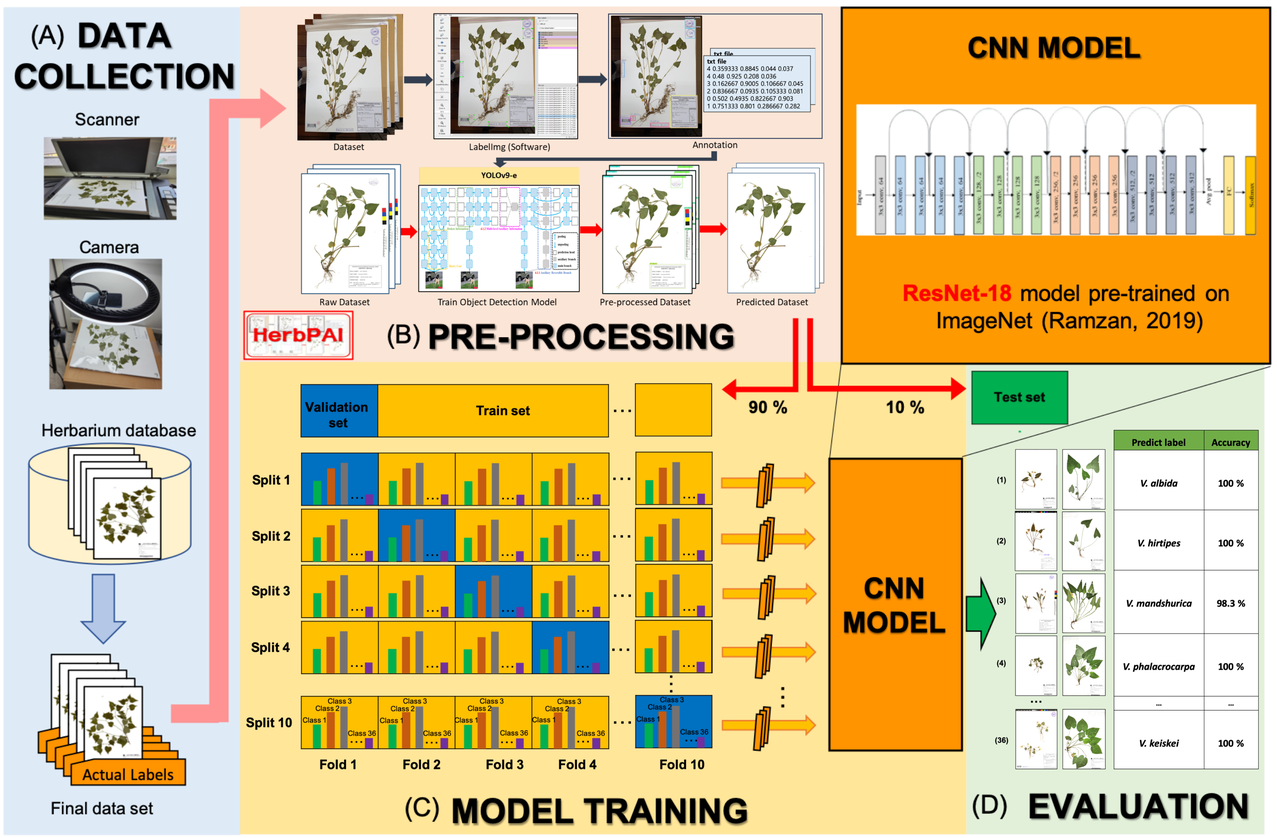

Figure 1. Schematic process showing the development of automatic herbarium image identification system based on CNN in this study.

Materials and Methods

Data collection and labeling (Figure 1. A)

- ***Data collection:***
  - Digitized herbarium images from the herbaria of **KH** (Lee and Yoo, 2020), **NIBR** (National Institute of Biological Resources), **SWU** (Sungshin Women's University), and **KWU** (Kangwon National University).
  - Web-based open image resources were added for some critical species: **iDigBio** (Integrated Digitized Biocollections) and **CVH** (Chinese Virtual Herbarium).
- ***Taxa labeling:***
  - Includes **36 Korean Violaceae** [ca. 40 taxa are native to Korea (Korea National Arboretum, 2020)] based on the availability of direct observation from the herbaria or web-based resources
  - Includes 31 ~ 600 images/taxon: **77.8 % (29 taxa) have > 300 images**.

Data preprocessing (Figure 2. B)

- ***Manual labeling of the training dataset***
  - Preparation of the training dataset (531 images)
  - Labeling the training dataset with the **labeling** program (https://github.com/HumanSignal/labelImg)
  - class > 0: non-plant specimen components (label, institution stamp, annotation label, barcode, palette, ruler, photo, envelope, map, tag, DB stamp and handwriting)
  - class = 0: specimen edges
- ***Automated labeling using YOLOv9***
  - Split the training dataset (Image set + Labelling set): 80% train set, 10% validation set, 10% test set
  - Training the YOLOv9 model: Obtain the weight file (best.pt) trained on the training dataset using the YOLOv9 model, which is a state-of-the-art object detection model.
  - Automated labeling of the entire dataset (14,939 images) with the weight files
- ***Batch removal of unnecessary information***
  - class > 0 (non-plant specimen components): cover with a white box
  - class = 0 (specimen edges): crop

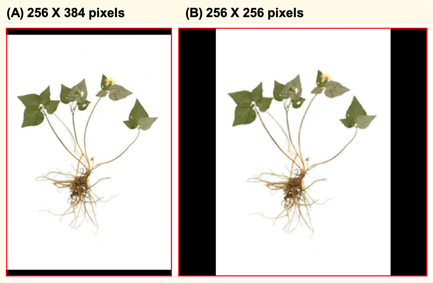

Figure 2. Visualization of resized input images for the AI studies. The black color indicates a zero-padding added to make a standard image size.

Model training (Figure 1. C)

- **Input image size tested**: **256 X 256** and **256 X 384 pixels with zero padding (Figure 2)**
- **Data splitting**: 90% train-validation set, 10% test set
- **CNN Model tested**: **ResNet-18** and **ResNet-34** (pre-trained models on ImageNet; Deng et al., 2009)
- **Adjustment of imbalanced dataset**: **stratified 10-fold cross-validation** to form folds while maintaining the proportion of images per species in the imbalanced dataset

Model performance evaluation (Figure 1. D)

- **Accuracy**: the ratio of correct predictions to the total number of predictions.
- **Precision**: the ratio of correct predictions for a specific class to the total number of predictions for that class.
- **Recall**: the ratio of correct predictions for a specific class to the total number of instances of that class.
- **F1-score**: the harmonic mean of precision and recall, providing a balanced measure of a model's performance, especially in imbalanced datasets
- **Used the macro average of all classes for each metric**: multi-class classification considered
- **Confusion matrix**: visualize the difference between the classes predicted by the model and the actual classes.

Results and Discussion

Evaluation of model performance (Table 2, Figure 3)

- Loss and accuracy curves suggest that our model was properly trained: the best classification performance was obtained for the **input image 256 x 384 pixels** and the **ResNet-18 model**, with **macro accuracy of 0.8651** and **macro F1-score of 0.7703**.

Confusion matrix (Figure 4)

- ***Can a classification model distinguish between two morphologically similar species?* (Figure 5 and 6)**
  - Out of 36 taxa, 29 taxa had classification accuracies above **97.9%**, while six taxa had low accuracies*: V. tokubuchiana* var. *takedana* (65.5%); *V. tricolor* (0%), *V. variegata* (0%), *V. violacea* (0%), *V. websteri* (0%), and *V. yazawana* (0%).
  - The six taxa with the highest percentage of wrong answers are considered to be very similar to taxa that can be misidentified by human vision. Exceptionally, the model could not classify *V. websteri*, even though it is clearly distinguishable by humans.
  - No correlation exists between the number of test set images per species and accuracy per species.
- ***Are classification models good at learning seasonal variation within species*? (Figure 5)**
  - **Importantly, accuracy was also high for taxa with large morphological changes between flowering and fruiting.**

Future studies

- We will improve our classification model by 1) increasing data for taxa with fewer than 50 images and 2) analyzing Grad-Cam data, a tool that shows where and how much each image data contributes to classification (**Figure 7** ; Shirai et al., 2022).
- We also plan to extend our studies 1) to **all tracheophytes in the Korean peninsula** through the national herbarium network and 2) to those in **East Asia** through international collaborations.

Table 2. Classification performance for different combinations of image sizes and models

Image size	Model	Macro accuracy	Macro F1-score	Macro precision	Macro recall
256 X 256	ResNet-18	0.8409	0.7419	0.7302	0.7766
	ResNet-34	0.8203	0.7225	0.7202	0.7561
256 X 384	ResNet-18	0.8651	0.7703	0.7578	0.8025
	ResNet-34	0.8572	0.7600	0.7487	0.7294

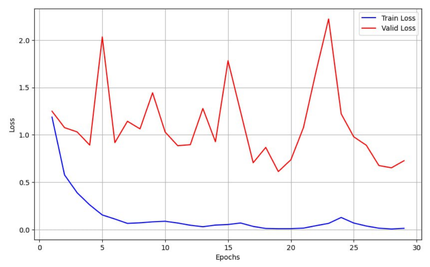

Figure 3. Loss curve graph during model training. Loss indicates the difference between the predicted class and the actual class.

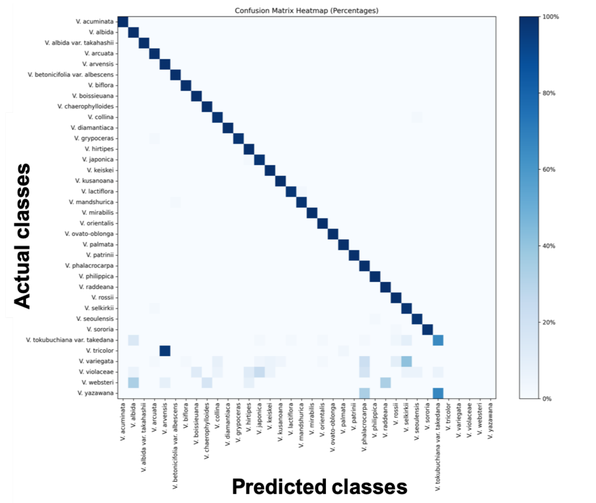

Figure 4. Confusion matrix.

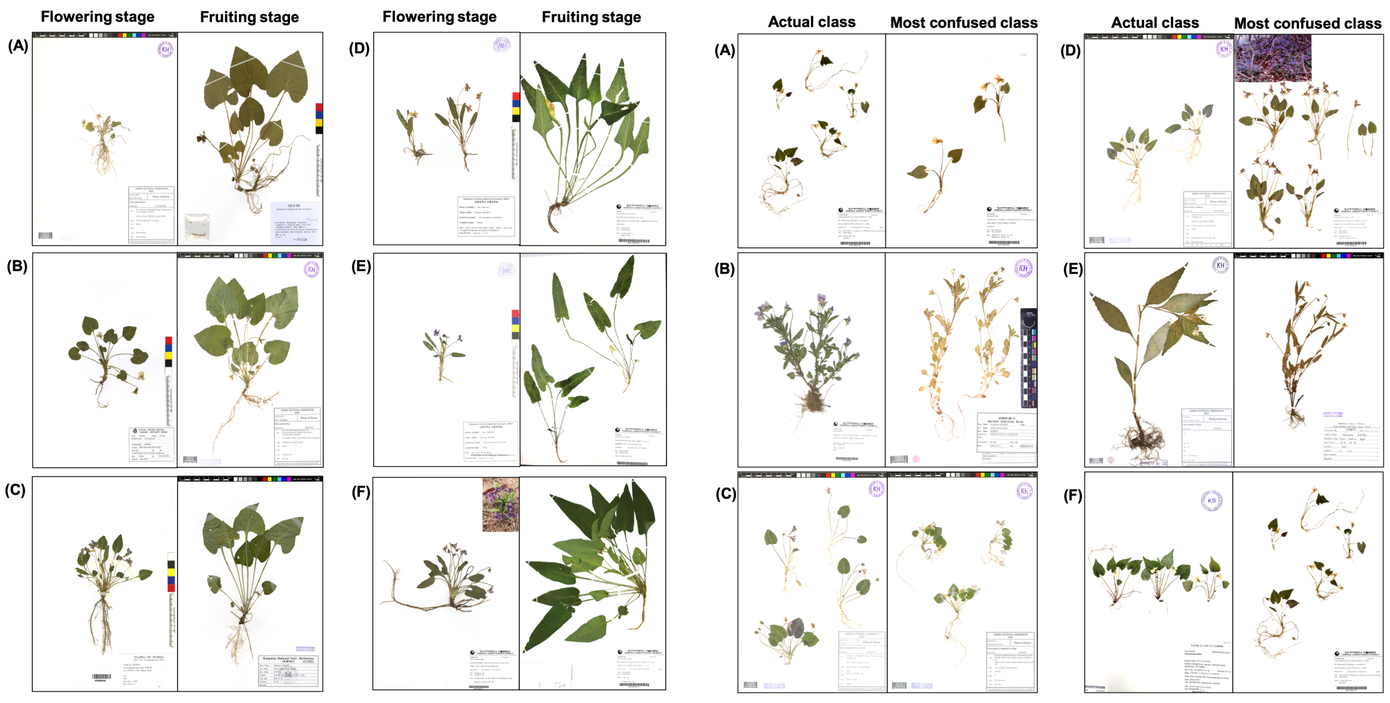

Figure 5. Images of representative specimens of taxa with a high degree of morphological variation in the different growing stages (season).
(A) *V. collina* (98.1%), (B) *V. keiskei* (100%), (C) *V. phalacrocarpa* (100%), (D) *V. mandshurica* (98.3%), (E) *V. philippica* (100%), and (F) *V. seoulensis* (97.9%); the numbers in parentheses indicate the classification accuracy of the model.

Figure 6. Example pairs of the six taxa that the model most often confused.
(A) *V. tokubuchiana* var. *takedana*: *V. albida* (14.5%), (B) *V. tricolor*: *V. arvensis* (96.8%), (C) *V. variegata*: *V. selkirkii* (41.2%), (D) *V. violacea*: *V. japonica* (23.5%), (E) *V. websteri*: *V. raddeana* (33.3%), and (F) *V. yazawana*: *V. tokubuchiana* var. *takedana* (66.7%); The number in parentheses indicates the percentage of false positives.

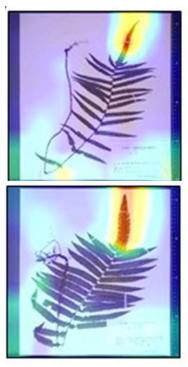

Figure 7. Example image of Grad-CAM analysis (Shirai et al., 2022).

Development of HerbPAI (Herbarium image Preprocessing for AI studies)

- HerbPAI **removes non-plant components** (labels, barcodes, stamps, scales, etc.) from herbarium images.
- HerbPAI was **developed with the YOLOv9 neural network** to preprocess herbarium images for AI studies.
- HerbPAI provides fast and reliable **preprocessing with high throughput** for large-scale herbarium image studies.

Implications

- HerbPAI improves accuracy in **AI classification studies on herbarium images** through image preprocessing
- HerbPAI contributes to **species conservation** by preventing the leakage of habitat information for endangered or rare plants when herbarium images are opened online.
- HerbPAI facilitates **international collaboration in AI studies** on multinational herbarium image data by excluding the textual information of species sensitive as biological resources (rare, medicinal, useful, etc.).

Download on Github: https://github.com/sujeong-han/Herb-PAI.

References

Cires Brown G, Guymer G, Franks A, Ranatunga D, Baba Y, Belongie SJ, Michelangeli FA, Ambrose BA, Little DP. 2022. The Herbarium 2021 Half-Earth Challenge Dataset and Machine Learning Competition. Frontiers in Plant Science. 12: 787127.

Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A. 2017. Going deeper in the automated identification of Herbarium specimens. BMC evolutionary biology 17.1: 1-14.

Davis CC. 2023. The herbarium of the future. Trends in Ecology & Evolution 38.5: 412-423.

de Lutio R, Park JY, Watson KA, D'Aronco S, Wegner JD, Wieringa JJ, Tulig M, Pyle RL, Gallaher TJ, Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition 248-255.

Goëau H, Bonnet P, Joly A. 2022. Overview of PlantCLEF 2022: Image-based plant identification at global scale. In CLEF 2022-Conference and Labs of the Evaluation Forum 3180.153: 1916-1928.

Hussein BR, Malik OA, Ong W-H, Slik JWF. 2022. Applications of computer vision and machine learning techniques for digitized herbarium specimens: a systematic literature review. Ecological Informatics 69: 101641.

Lee WT, Yoo KO. 2020. Violaceae Batsch. In: Park C, editor. Flora of Korea 4a. Incheon: National Institute of Biological Resources.

Shirai M, Takano A, Kurosawa T, Inoue M, Tagane S, Tanimoto T, Koganeyama T, Sato H, Terasawa T Horie T, Mandal I, Akihiro T. 2022. Development of a system for the automated identification of herbarium specimens with high accuracy. Scientific Reports. 12: 8066.

Acknowledgments

This work was supported by a grant from the National Institute of Biological Resource (**NIBRE202411**) and a grant from the National Research Foundation of Korea (**NRF-2017R1D1A1B03034952**).

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government (MSIT) (**2021-0-01341**, Artificial Intelligence Graduate School Program (Chung-Ang University)).

We thank the Korea National Arboretum (KH), National Institute of Biological Resources (NIBR), Kangwon National University (KWU), Integrated Digitized Biocollections (iDigBio), and Chinese Virtual Herbarium (CVH) for providing specimen images or permission to photograph them for our research.