

Crop Foliage Image Datasets

This file contains detailed information, including data collection, preprocessing, and data organization of the Foliagen datasets. The dataset can be accessed and downloaded from Hugging Face at the following link:

<https://huggingface.co/datasets/nabinpakka07/Foliagen>. The datasets are a collection of zip files as shown in figure below:



Dataset Name	Size	Format	Upload Method	Upload Date
soybean_bacterial_blight.zip	1.82 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_cercospora_leaf_blight.zip	1.8 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_downey_mildew.zip	1.77 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_frogeye.zip	1.83 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_healthy.zip	1.79 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_mosaic_virus.zip	1.78 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_potassium_deficiency.zip	1.8 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_rust.zip	1.82 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_sudden_death_syndrome.zip	1.86 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_target_spot.zip	1.78 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_test.zip	2.25 GB	LFS	Upload folder using huggingface_hub	11 days ago
soybean_val.zip	2.34 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_bacterial_spot.zip	1.72 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_early_blight.zip	1.74 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_healthy.zip	1.76 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_late_blight.zip	1.75 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_leaf_curl_virus.zip	1.76 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_leaf_mold.zip	1.75 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_mosaic_virus.zip	1.75 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_septoria_leaf_spot.zip	1.75 GB	LFS	Upload folder using huggingface_hub	11 days ago
tomato_target_spot.zip	1.76 GB	LFS	Upload folder using huggingface_hub	11 days ago

Data Collection

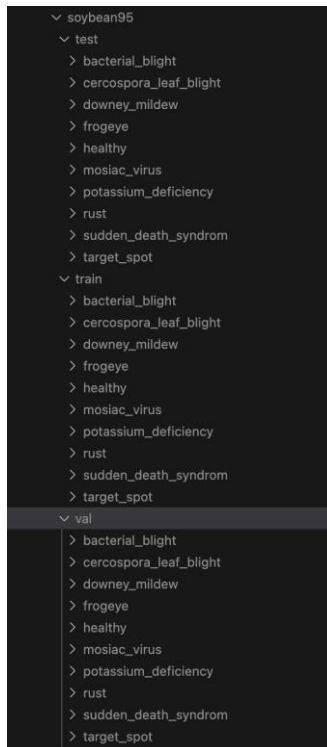
Soybean. A total of 10,722 high-resolution single leaf images covering 7 disease categories from ASDID [1] and 132 images (with 22 for Mosaic Virus and 110 for sudden Death Syndrome) from a Kaggle soybean diseased leaf dataset [2] were chosen for foliage image generation. Images of those nine disease categories feature diverse and complex natural/artificial backgrounds, which are undesirable when generating foliage images. Therefore, we pre-process the images to remove their undesired backgrounds (i.e., to make them transparent) using an open-sourced and AI enabled background remover, rembg [3]. Freely available soybean field soil images are then included in generated foliage images as their backgrounds, to emulate the natural habitat of soybean plants as best as possible.

The only publicly available and properly annotated dataset of diseased foliage images, MH-SoyaHealthVision [4], comprises two disease categories: soybean rust and mosaic virus. The original images have a very high resolution of 3840×2160 and the dataset is imbalanced, with rust images outnumbering healthy images by a factor of four, leading to biased predictions favoring the majority class and consequently reducing model's generalizability. To address these issues, we crop the high-resolution foliage images into ones with a lower

resolution to ensure a more balanced data distribution and to obtain a total of 3210 images, comprising 1084 rust images, 1027 healthy images, and 1099 mosaic virus images, respectively.

Tomato. The single-leaf diseased images of tomato are from the PlantVillage dataset [5], with 9 primary tomato disease categories and are taken in a laboratory environment and with the dimension of 256 × 256.

Data Organization



The datasets are split into 80% train, 10% validation, and 10% test dataset. The disease categories of validation and test datasets, due to their small size, are compressed into a single zip file. The naming convention of the validation and test datasets is ``<crop_name>_<test/val>.zip`` (tomato_val.zip). While training datasets, having large size (in GBs), are compressed separately into zip files according to the disease type. The naming convention for training datasets is ``<crop_name>_<disease_name>.zip``. For training purposes, the downloaded zip files are to be organized under 3 folders--train, test, and val--as illustrated in figure on the left.

References

1. Noah Bevers, Edward J Sikora, and Nate B Hardy. Soybean disease identification using original field images and transfer learning with convolutional neural networks. *Computers and Electronics in Agriculture*, 203:107449, 2022.
2. Sivism205. Soybean diseased leaf dataset, 2023.
3. Daniel Gatis. rembg: Remove image background. <https://github.com/danielgatis/rembg>, 2021. Accessed: 2025-05-08.

4. Sayali Shinde and Vahida Attar. MH-SoyaHealthVision: An indian UAV and leaf image dataset for integrated crop health assessment, 2024.
5. David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. arXiv preprint arXiv:1511.08060, 2015.