data

MDPI

Data Descriptor

LeafScans-Orchard: A Multi-Year Open RGB Scan Dataset of Orchard Plant Leaves for Species and Cultivar Classification

Paweł Chwietczuk [1], Seweryn Lipiński [1,*] and Paulina Chwietczuk [2]

[1] Faculty of Technical Sciences, University of Warmia and Mazury in Olsztyn, 10-036 Olsztyn, Poland; pawel.chwietczuk@uwm.edu.pl
[2] Independent Researcher, Różnowo, 11-001 Dywity, Poland; paulina.chwietczuk@gmail.com
* Correspondence: seweryn.lipinski@uwm.edu.pl

Abstract

LeafScans-Orchard is a curated, multi-year RGB image dataset of orchard plant leaves designed to support research in computer vision, machine learning, and plant phenotyping. The dataset comprises 9708 high-quality leaf scans acquired during collection campaigns conducted between 2015 and 2025, covering seven orchard crop species: apple, pear, sweet cherry, sour cherry, plum, peach, and apricot. In total, the dataset includes 67 cultivar labels. All samples were acquired using flatbed scanning under controlled conditions on a uniform background, ensuring high visual consistency and minimal background variability. The original scans were captured at 1200 dpi and subsequently converted into a public release format at 300 dpi, stored as lossless TIFF images to preserve morphological and textural details. Each image corresponds to a single leaf and is organized in a hierarchical directory structure by species, cultivar, and acquisition year, accompanied by image-level metadata and aggregated species–cultivar–year counts. LeafScans-Orchard is suitable for plant species classification, cultivar recognition, leaf morphology analysis, texture analysis, and general visual feature extraction. In addition to the main release, a representative subset of 300 original 1200 dpi scans is provided to support high-resolution analyses. The dataset is particularly suited for fine-grained classification, morphology-driven analysis, and methodological studies under controlled imaging conditions.

Dataset: LeafScans-Orchard Dataset (v1.0), Zenodo. DOI: https://doi.org/10.5281/zenodo.20187966 (accessed on 10 May 2026).

Dataset License: CC-BY 4.0

Keywords: leaf image dataset; orchard plants; leaf scans; RGB images; plant species classification; cultivar classification; agricultural datasets; computer vision; machine learning; image analysis

Check for updates

Academic Editor: Tianwen Zhang

Received: 20 May 2026
Revised: 19 June 2026
Accepted: 21 June 2026
Published: 23 June 2026

1. Summary

The increasing use of computer vision and machine learning methods in agriculture has highlighted the need for well-documented, high-quality image datasets acquired under controlled and reproducible conditions [1–6]. Leaf morphology remains an important source of visual information for plant identification, cultivar differentiation, and phenotypic analysis [7–10]. However, many publicly available leaf image datasets are limited in temporal coverage, label granularity, or relevance to fine-grained cultivar-level classification.

Existing public leaf image datasets have made important contributions to plant image analysis, including PlantVillage for plant disease recognition [11], Flavia [12] and Swedish Leaf [13,14] for leaf-based species classification, Leafsnap [15,16] for mobile plant identification, or MalayaKew [17,18] for leaf recognition under more standardized imaging conditions. These resources differ substantially in acquisition protocols, taxonomic scope, label structure, and intended applications. To the best of our knowledge, no widely available open leaf image dataset combines controlled image acquisition, orchard crop coverage, cultivar-level labels, and multi-year temporal structure within a single resource.

LeafScans-Orchard was prepared to address this gap by providing a curated collection of RGB flatbed scans of orchard plant leaves acquired under reproducible conditions. The dataset comprises 9708 images collected during acquisition campaigns conducted between 2015 and 2025 and covers seven economically important orchard crop species: apple, pear, sweet cherry, sour cherry, plum, peach, and apricot. In total, the dataset includes 67 cultivar labels.

Sample counts vary substantially across species, cultivars, and acquisition years, reflecting realistic long-term data collection conditions.

The main distinctive features of LeafScans-Orchard are its focus on orchard crops, cultivar-level label structure, multi-year acquisition period, controlled flatbed scanning protocol, lossless TIFF storage, and the inclusion of an additional high-resolution subset.

Images were acquired by controlled flatbed scanning on a uniform background. The public release consists of lossless 300 dpi TIFF images, complemented by a representative subset of 300 original 1200 dpi scans for high-resolution analyses.

The dataset is organized in a hierarchical directory structure by species, cultivar, and acquisition year, and is accompanied by structured metadata describing biological labels, acquisition information, file paths, and technical properties of each image. A summary of the dataset is provided in Table 1.

Table 1. Overview of the LeafScans-Orchard dataset.

Attribute	Value
Number of images	9708
Species	7 orchard crop species
Cultivars	67 cultivar labels
Acquisition years	2015, 2022, 2023, 2024, 2025
Image type	RGB leaf scans
Acquisition method	Flatbed scanning
Background	Uniform
Original resolution	1200 dpi
Public release resolution	300 dpi
High-resolution subset	300 images at 1200 dpi
File format	TIFF
Compression	Lossless
Availability	Zenodo
License	CC BY 4.0

The dataset has been prepared for open dissemination via Zenodo, as it is intended to serve as a reusable resource for computer vision, machine learning, plant phenotyping, and methodological studies focused on leaf-based visual analysis in orchard crops.

2. Data Description

2.1. Dataset Overview

The LeafScans-Orchard dataset comprises 9708 RGB images of orchard plant leaves acquired during five collection years between 2015 and 2025. The dataset covers seven orchard crop species and includes 67 cultivar-level labels (Table 1). All images were obtained using controlled flatbed scanning and released in a standardized format to support reproducible analysis and reuse.

The main public release consists of lossless 300 dpi TIFF images. It is complemented by a representative high-resolution subset of 300 images provided at the original acquisition resolution of 1200 dpi. These two components support both large-scale computational experiments and analyses requiring access to finer spatial detail.

The dataset is organized by species, cultivar, and acquisition year. It is accompanied by image-level metadata and aggregated species–cultivar–year counts, which document the biological labels, acquisition information, file paths, and class distribution of the dataset. Detailed information on image format, high-resolution subset selection, and metadata structure is provided in the following sections.

2.2. Species and Cultivar Composition

The LeafScans-Orchard dataset includes seven orchard crop species: sweet cherry, apple, sour cherry, plum, pear, peach, and apricot. In total, 67 cultivar labels are represented, including a small number of samples labeled as unidentified at the cultivar level in the full 300 dpi release. The number of cultivars and images varies substantially between species, reflecting differences in sampling availability, agricultural relevance, and the multi-year character of data collection.

Sweet cherry constitutes the largest portion of the dataset, with 5491 images distributed across eight cultivar labels. Several sweet cherry cultivars, including Burlat, Kordia, Van, and Vega, are represented by more than one thousand images each, whereas less frequent cultivars contain only several dozen images or fewer. This structure enables both species-level classification and fine-grained cultivar-level discrimination within a single orchard species.

Apple is the most diverse species in terms of cultivar representation, with 31 cultivar labels and 1368 images in total. Most apple cultivars are represented by relatively small numbers of samples, whereas a limited number of cultivars, such as Ligol and Novamac, contain several hundred images. This pronounced within-species imbalance provides a useful test case for evaluating classification methods under uneven cultivar-level sample distributions.

Sour cherry and plum are represented by 1199 and 767 images, respectively, with several cultivars per species. These groups provide additional material for comparative analyses of leaf shape, venation, and texture across closely related cultivars. Pear includes 745 images distributed across nine cultivar labels, including both moderately represented and rare cultivars.

Peach is represented by two cultivar labels, whereas apricot is represented by a single cultivar label. These species therefore contribute primarily to species-level classification and intra-cultivar variability analyses, with only limited cultivar-level analysis possible for peach.

Overall, the dataset exhibits a strongly imbalanced distribution at the cultivar level, with sample counts ranging from single-digit values to more than one thousand images per cultivar. This distribution was not artificially balanced. Instead, it reflects realistic long-term collection conditions and supports research on class imbalance, fine-grained visual discrimination, and robust model evaluation across heterogeneous biological categories.

Table 2 summarizes the dataset composition at the species level, while Figure 1 shows the distribution of image counts across individual cultivar labels.

Table 2. Aggregated composition of the LeafScans-Orchard dataset by species.

Species	Number of Cultivars	Number of Images
Sweet cherry	8	5491
Apple	31	1368
Sour cherry	4	1199
Plum	12	767
Pear	9	745
Peach	2	64
Apricot	1	74
Total	67	9708

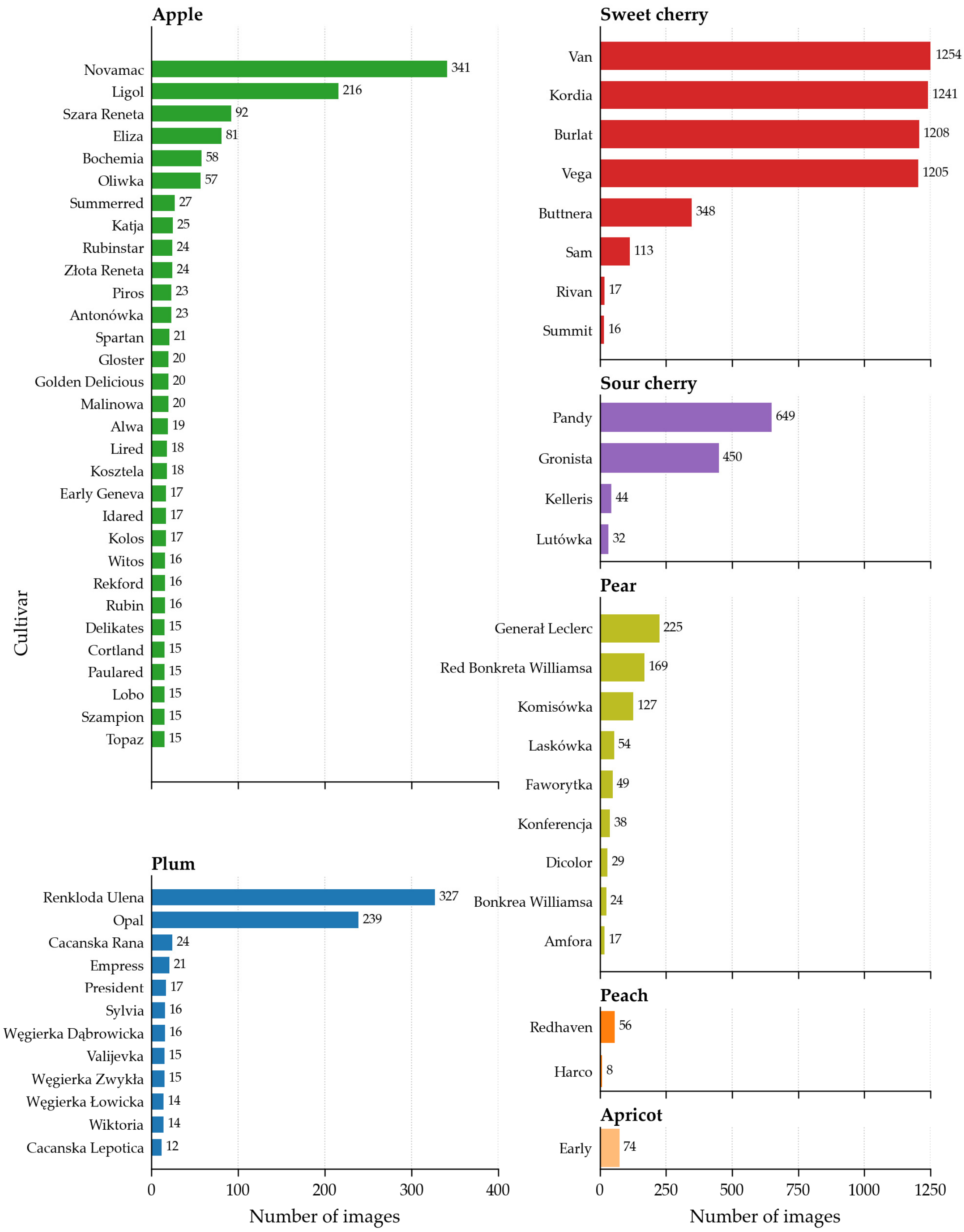

Figure 1. Distribution of image counts across cultivars in the LeafScans-Orchard dataset.

2.3. Temporal Coverage

The LeafScans-Orchard dataset spans a multi-year acquisition period between 2015 and 2025 and includes five distinct collection years. This temporal structure introduces natural variability in both the number of samples and the set of cultivars represented in individual acquisition years. Such variability reflects realistic data collection conditions rather than a fixed, uniformly balanced sampling protocol.

The number of images collected varies substantially between acquisition years. Some collection campaigns contribute a large proportion of the dataset, whereas others contain more limited numbers of samples. Similarly, the presence and abundance of individual cultivars vary across years. Some cultivars are represented in multiple acquisition years, while others appear only in a single year or sporadically across the dataset. As a result, the temporal dimension of LeafScans-Orchard is characterized by non-uniform sample distribution and shifting cultivar composition.

This multi-year structure enables longitudinal analyses and the evaluation of model robustness with respect to temporal variability, including changes in class distribution over time.

2.4. Image Acquisition and Format

Images in the LeafScans-Orchard dataset were acquired under controlled conditions using two flatbed scanners: a Canon CanoScan LiDE 30 (Canon Inc., Ōta, Tokyo, Japan) for the 2015 scans and a Plustek OpticSlim 1180 (Plustek Inc., Taipei, Taiwan) for the scans acquired in 2022–2025.

Although both devices were used under controlled conditions and at the same nominal optical resolution, scanner-specific differences may affect color response, illumination characteristics, local contrast, and fine texture representation. No retrospective color calibration target or scanner cross-calibration procedure was applied. Therefore, scanner-related variability should be considered a potential source of technical variation, particularly in longitudinal analyses comparing the 2015 scans with scans acquired in 2022–2025.

Individual leaves were placed on a uniform background prior to scanning, which minimized background interference and provided a consistent visual context across samples. This acquisition setup was selected to preserve the natural shape and surface characteristics of the leaves while reducing variability caused by uncontrolled illumination or environmental conditions.

The scans were captured as RGB images at an original optical resolution of 1200 dpi. This resolution preserves fine morphological and textural details and serves as the reference acquisition quality for the dataset. For the primary public release, all images were uniformly converted to 300 dpi, providing a practical balance between file size and information content while maintaining compatibility with common image analysis and machine learning workflows.

All released images are stored in TIFF format with lossless compression. The file format and resolution are consistent across the public dataset release. No additional color normalization, geometric transformation, segmentation, augmentation, or image enhancement was applied beyond resolution conversion. Each file corresponds to a single leaf and is uniquely identifiable through a consistent naming scheme and the accompanying metadata.

Representative examples of leaf scans from selected orchard crop species are shown in Figure 2. The examples illustrate the uniform black background, controlled flatbed scanning conditions, and visible interspecific differences in leaf shape, size, and surface appearance.

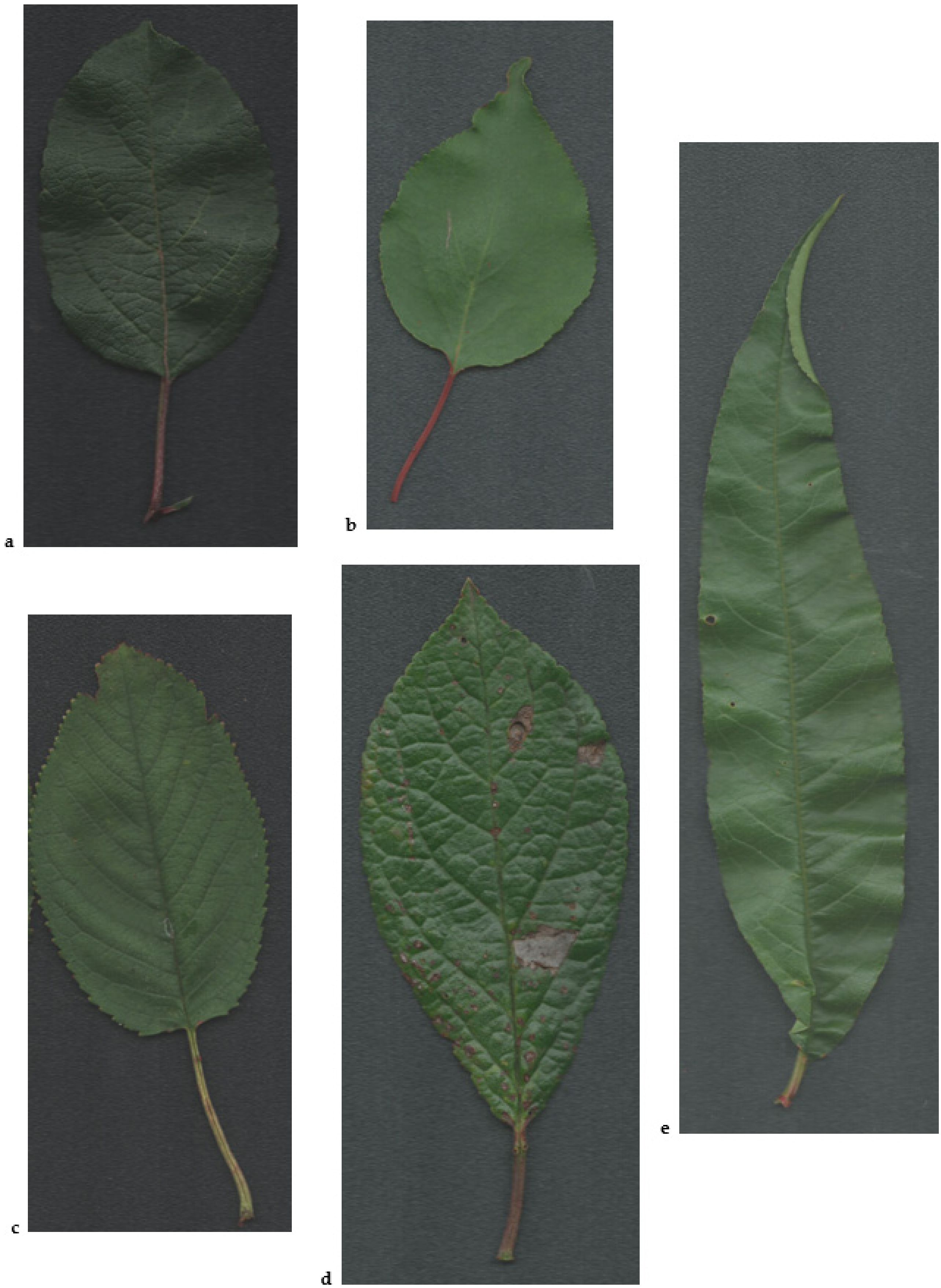

Figure 2. Representative leaf scans from selected orchard crop species in the LeafScans-Orchard dataset: (**a**) apple, (**b**) apricot, (**c**) sweet cherry, (**d**) plum, and (**e**) peach.

2.5. High-Resolution Subset

In addition to the main public release at 300 dpi, the LeafScans-Orchard dataset includes a high-resolution subset of images provided at the original scanning resolution of 1200 dpi. This subset is intended to support analyses requiring maximum spatial detail, such as scale-sensitive feature extraction, texture analysis, contour-based measurements, and resolution-dependent performance evaluation.

The high-resolution subset comprises 300 images selected from the full dataset with an emphasis on maximizing cultivar diversity. All samples included in this subset are fully identified at both species and cultivar levels. Dominant cultivars are represented by multiple images to capture intra-cultivar variability, while rarer cultivars are included whenever possible to ensure broad coverage of morphological diversity.

The file organization and naming conventions of the high-resolution subset mirror those of the full 300 dpi release, enabling direct comparison between resolutions. For each image included in the subset, the 300 dpi and 1200 dpi versions share the same unique identifier and corresponding metadata entries. This design facilitates reproducible experiments that explicitly investigate the impact of image resolution on analytical outcomes.

The high-resolution subset complements the primary dataset release by providing reference-quality images without substantially increasing the overall data volume. It therefore balances accessibility with the preservation of fine structural information. The composition of the high-resolution subset is given in Table 3.

Table 3. Composition of the high-resolution 1200 dpi subset of the LeafScans-Orchard dataset.

Species	Number of Cultivars	Number of Images	Notes
Sweet cherry	8	100	Dominant and rare cultivars included
Apple	30	40	Emphasis on cultivar diversity
Sour cherry	4	35	Balanced representation of major cultivars
Plum	12	40	High morphological variability
Pear	8	35	Frequent and rare cultivars included
Peach	2	30	Two cultivars
Apricot	1	20	Single cultivar
Total	—	300	Fully identified samples only

The complete archive of original 1200 dpi scans is retained locally by the authors. Owing to its substantially larger data volume, it is not included in the current public release. The publicly available 300-image subset was selected to provide representative access to the original acquisition quality and to support resolution-dependent analyses. The possibility of extending the publicly available high-resolution material may be considered in future dataset versions.

2.6. Metadata and File Organization

The metadata file was designed to support FAIR reuse and to remain broadly compatible with the logic of plant phenotyping metadata standards such as MIAPPE [19,20], while preserving a lightweight structure appropriate for an image dataset. Relation of metadata fields included in LeafScans-Orchard to MIAPPE/FAIR descriptors is given in Appendix A (Table A1).

The LeafScans-Orchard dataset is distributed using a hierarchical file organization scheme designed to facilitate both manual inspection and automated processing. Image files are arranged by biological label and acquisition year. At the top level, directories correspond to species, followed by cultivar-specific subdirectories and year-specific folders.

Each image file represents a single leaf and is uniquely identifiable through a consistent file naming convention.

The dataset is accompanied by four machine-readable CSV and JSON files. The first two files, LeafScans-Orchard_metadata.csv and LeafScans-Orchard_metadata.json, provide image-level metadata, with each row in the CSV file and each record in the JSON file corresponding to one image. These files include the unique image identifier, species name, scientific name, cultivar label, cultivar identification status, acquisition year, acquisition date, collection location code, image resolution, file format, and relative file path within the dataset archive. For images included in the high-resolution subset, the metadata also indicate the availability of the corresponding 1200 dpi version and provide the relevant file path where applicable.

The remaining two files, LeafScans-Orchard_summary_counts.csv and LeafScans-Orchard_summary_counts.json, provide aggregated counts by species, cultivar, and acquisition year. These files summarize the class distribution and temporal structure of the dataset and enable users to verify the number of images available for each species–cultivar–year combination. They are intended as a compact overview of the dataset composition, whereas the image-level metadata files should be treated as the authoritative source for selecting, filtering, and analyzing individual images.

File names follow a consistent and descriptive naming convention encoding cultivar identity, acquisition date, sampling location, and image sequence number. Each file name consists of the cultivar name, acquisition date in the YYYY_MM_DD format, a single-letter location identifier, and a sequential image number.

All collection sites are in the Warmian–Masurian Voivodeship in north-eastern Poland. The location codes provide site-level geographic information.

The location identifier provides coarse-grained information on the place of leaf collection: B denotes Bielice, K denotes Kronowo, O denotes Olsztyn, R denotes Różnowo location 1, R1 denotes Różnowo location 2 and S denotes Sętal.

For example, the file

Sweet Cherry (Prunus avium (L.) L.)/Burlat/2025/burlat_2025_09_01_R_121.tif represents the scan of leaf no. 121 of the cultivar Burlat, collected in Różnowo on 1 September 2025.

For reproducibility and downstream processing, users are encouraged to rely on LeafScans-Orchard_metadata.csv or LeafScans-Orchard_metadata.json rather than directory names alone. The summary count file can be used to inspect the overall dataset structure, class imbalance, and temporal distribution of samples before designing classification, morphology, or machine-learning experiments.

3. Methods

3.1. Data Collection

Leaves were collected during the vegetation season, mainly between June and October, depending on species, orchard accessibility, and harvest-related constraints. In stone fruit species, especially sweet cherry and sour cherry, leaf collection was often performed after fruit harvest to avoid interfering with orchard production. Therefore, the dataset does not represent early spring leaf development, but includes leaves collected during later vegetative stages, including both younger leaves formed on current-year shoots and older fully developed leaves.

Where possible, leaves were collected from more than one tree per cultivar. In Sętal (location identifier—S) three to five trees per cultivar were available. However, tree-level identifiers were not recorded systematically, and the dataset should therefore not be interpreted as a tree-replicated phenotyping experiment. The sampling unit in the released dataset is the individual scanned leaf, with cultivar identity assigned according to orchard documentation available at the time of collection.

All images included in the LeafScans-Orchard dataset were acquired using flatbed scanning under controlled conditions. Individual leaves were collected from orchard plants and placed on a uniform background prior to scanning. This setup was selected to minimize background variability and illumination-related artifacts while preserving the natural shape and surface characteristics of the leaves.

The scans were captured as RGB images at an original optical resolution of 1200 dpi. Scanning parameters were kept consistent within each acquisition campaign to ensure visual comparability across samples collected in the same period. Data collection was carried out during five acquisition years between 2015 and 2025, resulting in a dataset with substantial inter-annual variability in species composition, cultivar representation, and sample counts.

The same general sampling principle was followed across acquisition years. Leaves suitable for flatbed scanning were collected during the vegetation season, while the exact collection date, number of available trees, and proportions of younger and older leaves varied according to species, orchard accessibility, and harvest timing.

Each scan corresponds to a single leaf. No artificial augmentation, color normalization, geometric transformation, or segmentation was applied during data acquisition. The dataset therefore preserves the original visual appearance of the scanned leaves within the constraints of the controlled flatbed scanning setup.

3.2. Data Processing

The original 1200 dpi scans were archived as uncompressed raster images and subsequently processed to generate the public dataset release. For distribution and practical reuse, all images were converted to 300 dpi and stored as TIFF files with lossless compression. This resolution was selected as a compromise between file size and preservation of morphological and textural detail.

The downsampling process was applied uniformly across all images using IrfanView 4.73 [21]. No additional cropping beyond the original scan boundaries was performed, and the orientation of each leaf was preserved. The RGB color channels were retained without additional correction, enhancement, or normalization.

After conversion, image files were organized into a hierarchical directory structure by species, cultivar, and acquisition year. File names were standardized according to a unique identifier scheme encoding cultivar identity, acquisition date, collection location, and image sequence number. Image-level labels and technical attributes were recorded in LeafScans-Orchard_metadata.csv, while aggregated species–cultivar–year counts were summarized in LeafScans-Orchard_summary_counts.csv.

3.3. High-Resolution Subset Selection

In addition to the main 300 dpi release, a high-resolution subset of 300 images was prepared at the original scanning resolution of 1200 dpi. This subset was designed to support analyses requiring maximum spatial detail, such as scale-sensitive feature extraction, texture analysis, contour-based measurements, and resolution-dependent performance evaluation.

The subset was selected with an emphasis on maximizing cultivar diversity. Only samples with fully identified species and cultivar labels were included. Dominant cultivars were represented by multiple samples to capture intra-cultivar variability, while rarer cultivars were included whenever possible to broaden morphological coverage.

The file structure and naming conventions of the high-resolution subset mirror those of the full 300 dpi release. For each image included in the subset, the corresponding 300 dpi and 1200 dpi versions share the same unique identifier, allowing direct comparison between resolutions and reproducible evaluation of resolution-dependent effects.

3.4. Data Quality Control and Limitations

Data curation was performed to ensure structural consistency, label traceability, and technical usability of the dataset. The number of image files was checked against the expected counts for each species, cultivar, and acquisition year. File names, directory structure, and metadata entries were verified for consistency, and the summary count table was used to confirm the reported class distribution. The presence of corresponding metadata entries for the released images was also checked.

Cultivar labels were assigned based on official orchard documentation provided by the plantation operators at the time of leaf collection.

Exact tree-level geographic coordinates and detailed environmental descriptors, including local weather conditions, soil properties, orchard-management practices, and microclimatic variables, were not recorded systematically and are therefore not included in the dataset.

All released images were visually inspected during curation to confirm file readability, appropriate placement of leaves on the background, and the absence of obvious file corruption or severe scanning artifacts. No automated quantitative image-quality criteria, such as blur scores, color-calibration metrics, or artifact-detection thresholds, were applied. Images with evident technical problems were excluded during curation; however, no separate exclusion log was retained.

The controlled scanning setup provides high visual consistency and low background noise across the dataset. However, natural biological variability in leaf size, shape, venation, surface texture, and visual condition was intentionally preserved and constitutes an inherent characteristic of the data.

Leaves were not selected to represent only ideal or disease-free material. The dataset intentionally preserves natural variation in leaf appearance, including differences in age, size, shape, color, surface texture, minor mechanical damage, pest-related marks, disease symptoms, and weather-related damage. Leaves that were strongly curled, severely deformed, or unsuitable for reliable flatbed scanning were excluded during initial curation, because they could not be positioned sufficiently flat on the scanner surface. The high-resolution 1200 dpi subset was selected to contain technically clear scans without major damage, strong folding, or pronounced scanning artifacts.

The dataset has several limitations that should be considered when designing experiments. First, the naturally imbalanced class structure described in Section 2.2 may affect model training, validation, and interpretation of classification results. The dataset also does not include pixel-level segmentation masks, disease labels, detailed environmental metadata, or field-acquired photographs. Consequently, tasks such as semantic segmentation, disease detection, or direct deployment in field conditions would require additional annotation, external data, or domain adaptation.

4. Potential Applications

The LeafScans-Orchard dataset is designed to support a broad range of applications in computer vision, machine learning, plant phenotyping, and agricultural image analysis. Owing to its controlled acquisition conditions, uniform background, and consistent flatbed scanning protocol, the dataset provides high-quality visual material suitable for both classical image analysis and modern learning-based approaches.

https://doi.org/10.3390/data11070153

One of the primary applications of the dataset is plant species classification. The presence of seven orchard crop species and 9708 images enables the development and evaluation of models for leaf-based species recognition under controlled imaging conditions. Equally important is the cultivar-level structure of the dataset. The inclusion of 67 cultivar labels, with highly variable sample counts, supports fine-grained visual classification and the discrimination of closely related cultivars based on subtle differences in leaf shape, venation, contour, texture, and surface appearance.

The naturally imbalanced cultivar-level structure makes the dataset particularly suitable for studying long-tailed and imbalanced classification. Several cultivars are represented by large sample sets, whereas others contain only a small number of images. This enables the evaluation of strategies such as class-weighted learning, resampling, focal-loss-based training, few-shot recognition, and hierarchical classification approaches in which species-level recognition is separated from cultivar-level discrimination. The dataset can also be used to investigate how class imbalance affects model calibration, minority-class sensitivity, and the reliability of cultivar-level predictions.

The controlled flatbed scanning protocol makes LeafScans-Orchard suitable for reproducible studies of leaf morphology and texture. Potential applications include segmentation-free feature extraction, contour-based descriptors, shape indices, venation-related characteristics, and texture features. Because all images were acquired on a uniform background and released in lossless TIFF format, the dataset can also serve as a benchmark for comparing classical feature engineering with convolutional neural networks, vision transformers, or hybrid approaches combining handcrafted and learned descriptors.

The multi-year acquisition structure enables longitudinal and cross-year validation scenarios. For example, models may be trained on selected acquisition years and tested on later years to evaluate temporal robustness, dataset shift, and changes in cultivar composition over time. Because the 2015 scans and the 2022–2025 scans were acquired using different scanner models, the dataset may also support studies on scanner- or acquisition-period-related domain shift, if scanner model and acquisition year are taken into account during experimental design.

The high-resolution subset further enables resolution-dependent studies. Users may compare results obtained from the standard 300 dpi release and the original 1200 dpi scans to assess the influence of image resolution on morphology extraction, texture analysis, fine-structure representation, and classification performance. This may be useful for determining whether high-resolution scanning provides measurable benefits for cultivar-level recognition or whether lower-resolution images are sufficient for specific analytical tasks.

Finally, LeafScans-Orchard can be used as a reference dataset for educational and methodological studies focused on reproducible plant image analysis, benchmark construction, class imbalance handling, year-aware validation, and the interpretation of biological variability in image datasets.

5. User Notes

Users of the LeafScans-Orchard dataset should consider several practical aspects when working with the data. The dataset was collected under controlled conditions using flatbed scanning and a uniform background. As a result, the images are characterized by low background variability and high visual consistency, which may differ substantially from field-acquired photographs. Models trained exclusively on this dataset may therefore require additional adaptation before being applied to in situ images.

Users should also consider the acquisition year and scanner model when designing longitudinal or cross-year analyses. Because the 2015 scans and the 2022–2025 scans were acquired using different scanners (model can be inferred from acquisition year), models

https://doi.org/10.3390/data11070153

trained across acquisition years may capture both biological variation and scanner-related technical variation. For this reason, year-aware validation, scanner-aware train/test splits, or explicit domain-shift evaluation are recommended when assessing model robustness across acquisition periods. When color-sensitive features are used, users may additionally consider applying color normalization or evaluating whether results remain stable across acquisition years.

The dataset contains a strongly imbalanced distribution of cultivar labels, with some cultivars represented by more than one thousand images and others by only a few samples. This imbalance reflects realistic long-term data availability rather than curated class balancing. Users are encouraged to account for this characteristic when designing experiments, for example by applying stratified sampling, class-balanced evaluation metrics, cross-validation strategies, or separate species-level and cultivar-level evaluation protocols.

Although all images are labeled at the species level, a small subset of samples in the full 300 dpi release is labeled as unidentified at the cultivar level. These samples may be excluded from cultivar-level classification tasks if strict cultivar label consistency is required. For analyses requiring unambiguous cultivar labels or maximum spatial detail, users may use the provided high-resolution 1200 dpi subset, which contains only fully identified samples.

No pixel-level annotations, segmentation masks, disease labels, detailed environmental metadata, or field-acquired images are included in the dataset. Consequently, tasks such as semantic segmentation, disease detection, or direct field deployment require additional annotation, external data sources, or domain adaptation. Users interested in texture- or scale-sensitive analyses may benefit from comparing results obtained using the 300 dpi release and the 1200 dpi subset.

The dataset is provided in a hierarchical directory structure and is accompanied by image-level metadata and summary count files in both CSV and JSON formats.

Any derived datasets, annotations, trained models, or published analyses should clearly reference the specific dataset version and image resolution used.

Author Contributions: Conceptualization, P.C. (Paweł Chwietczuk) and S.L.; methodology, P.C. (Paweł Chwietczuk) and S.L.; formal analysis, P.C. (Paweł Chwietczuk) and S.L.; investigation, P.C. (Paweł Chwietczuk) and P.C. (Paulina Chwietczuk); data curation, P.C. (Paweł Chwietczuk) and P.C. (Paulina Chwietczuk); writing—original draft preparation, S.L. and P.C. (Paweł Chwietczuk); writing—review and editing, P.C. (Paulina Chwietczuk); visualization, P.C. (Paweł Chwietczuk) and S.L.; supervision, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset described in this article is openly available in Zenodo as LeafScans-Orchard Dataset (v1.0.0) at https://doi.org/10.5281/zenodo.20187966 (accessed on 10 May 2026). The repository includes the 300 dpi image release, the 1200 dpi high-resolution subset, image-level metadata, aggregated species–cultivar–year counts, and supporting documentation. The complete archive of original 1200 dpi scans is retained locally by the authors but is not included in the current public release because of its substantially larger data volume.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Metadata fields included in LeafScans-Orchard and their relation to MIAPPE/FAIR descriptors.

Metadata Field	Description	MIAPPE/FAIR Relevance
image_id	Unique image identifier	Sample/data object identifier
species	Common species name	Biological material/taxon
scientific_name	Scientific species name	Taxonomic identification
cultivar	Cultivar label	Biological material/genotype descriptor
cultivar_status	Cultivar identification status	Data quality/label status
acquisition_year	Year of acquisition	Observation/acquisition time
acquisition_date	Date of acquisition	Time point
location_code	Collection site code	Sampling location
resolution_dpi	Image resolution	Technical imaging metadata
file_format	Image file format	Data format
relative_path	Path to image file	Data access/interoperability

References

1. Caroppo, A.; Giovanni, D.; Leone, A. A Systematic Review of Available Multispectral UAV Image Datasets for Precision Agriculture Applications. *Remote Sens.* **2026**, *18*, 659. [CrossRef]
2. Gálvez-Salido, A.; Robles, F.; Gonçalves, R.J.; de la Herrán, R.; Ruiz Rejón, C.; Navajas-Pérez, R. Analysis of Biological Images and Quantitative Monitoring Using Deep Learning and Computer Vision. *J. Imaging* **2026**, *12*, 88. [CrossRef] [PubMed]
3. Liao, B.; Wang, Y.; Li, X.; Cheng, Y.; Zhao, G.; Zhou, Y. Deep Learning-Based Image Recognition for Food Science and Technology: End-to-End Workflows and Domain-Specific Solutions. *Compr. Rev. Food Sci. Food Saf.* **2026**, *25*, e70388. [CrossRef] [PubMed]
4. Lipiński, S. Statistical Analysis of Seeds Morphology and Texture for Interspecific Similarity Assessment across Taxonomic Levels. *Biul. Inst. Hod. Aklim. Roślin* **2025**, *303/304*, 19–33. [CrossRef]
5. Murphy, K.M.; Ludwig, E.; Gutierrez, J.; Gehan, M.A. Deep Learning in Image-Based Plant Phenotyping. *Annu. Rev. Plant Biol.* **2024**, *75*, 771–795. [CrossRef] [PubMed]
6. Zhu, L.; Tang, L.; Ren, S. Deep Learning-Based Text Generation for Plant Phenotyping and Precision Agriculture. *Front. Plant Sci.* **2025**, *16*, 1564394. [CrossRef] [PubMed]
7. Nasiri, A.; Taheri-Garavand, A.; Fanourakis, D.; Zhang, Y.-D.; Nikoloudakis, N. Automated Grapevine Cultivar Identification via Leaf Imaging and Deep Convolutional Neural Networks: A Proof-of-Concept Study Employing Primary Iranian Varieties. *Plants* **2021**, *10*, 1628. [PubMed]
8. Saleem, G.; Akhtar, M.; Ahmed, N.; Qureshi, W.S. Automated Analysis of Visual Leaf Shape Features for Plant Classification. *Comput. Electron. Agric.* **2019**, *157*, 270–280. [CrossRef]
9. Wang, Y.-H.; Su, W.-H. Convolutional Neural Networks in Computer Vision for Grain Crop Phenotyping: A Review. *Agronomy* **2022**, *12*, 2659. [CrossRef]
10. Wen, W.; Wang, J.; Zhao, Y.; Wang, C.; Liu, K.; Chen, B.; Wang, Y.; Duan, M.; Guo, X. 3D Morphological Feature Quantification and Analysis of Corn Leaves. *Plant Phenomics* **2024**, *6*, 0225. [CrossRef] [PubMed]
11. Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using Deep Learning for Image-Based Plant Disease Detection. *Front. Plant Sci.* **2016**, *7*, 1419. [CrossRef] [PubMed]
12. Flavia Dataset. Available online: https://www.kaggle.com/datasets/gauravneupane/flavia-dataset (accessed on 10 May 2026).
13. Swedish Leaf Dataset. Available online: https://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/ (accessed on 10 May 2026).
14. Söderkvist, O. Computer Vision Classification of Leaves from Swedish Trees. Master's Thesis, Linköping University, Linköping, Sweden, 2001.
15. LeafSnap. Available online: https://leafsnap.com/ (accessed on 10 May 2026).
16. Kumar, N.; Belhumeur, P.N.; Biswas, A.; Jacobs, D.W.; Kress, W.J.; Lopez, I.C.; Soares, J.V.B. Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In *Computer Vision—ECCV 2012*; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7573, pp. 502–516. ISBN 978-3-642-33708-6.
17. MalayaKew Plant Leaf Dataset. Available online: https://www.kaggle.com/datasets/abdulhasibuddin/malayakew-plant-leaf-dataset (accessed on 10 May 2026).

18. Lee, S.H.; Chan, C.S.; Wilkin, P.; Remagnino, P. Deep-Plant: Plant Identification with Convolutional Neural Networks. In *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2015; pp. 452–456.
19. Miappe. Available online: https://www.miappe.org/ (accessed on 10 May 2026).
20. Papoutsoglou, E.A.; Faria, D.; Arend, D.; Arnaud, E.; Athanasiadis, I.N.; Chaves, I.; Coppens, F.; Cornut, G.; Costa, B.V.; Ćwiek-Kupczyńska, H.; et al. Enabling Reusability of Plant Phenomic Datasets with MIAPPE 1.1. *New Phytol.* **2020**, *227*, 260–273. [CrossRef] [PubMed]
21. IrfanView Homepage. Available online: https://www.irfanview.com/ (accessed on 10 May 2026).