Received: 10 June 2025 | Revised: 15 October 2025 | Accepted: 4 November 2025

DOI: 10.1002/ppp3.70149

METHODS AND TECHNIQUES

Plants People Planet PPP Open Access

A pipeline to compile expert-verified datasets of digitised herbarium specimens for automated plant identification to accelerate taxonomy

Jed Arno[1] | Jérémie Morel[2,3] | Fitiavana Rasaminirina[3,4,5] | Juliene de Fátima Maciel-Silva[3,6,7] | Daniel Cahen[3] | Damon P. Little[8] | Daniele Silvestro[9,10] | Alexandre Antonelli[3,10,11] | Olwen Grace[3,12] | Li Zhang[1] | Isabel Larridon[3]

[1]Department of Computer Science, Royal Holloway, University of London, Surrey, UK

[2]Department of Life Sciences, Imperial College, Silwood Park Campus, Ascot, UK

[3]Royal Botanic Gardens, Kew, Richmond, UK

[4]Department of Plant Biology and Ecology, University of Antananarivo, Antananarivo, Madagascar

[5]Royal Botanic Gardens, Kew, Kew Madagascar, Antananarivo, Madagascar

[6]Inst. Ciências Biológicas, Universidade Federal do Pará, Belém, Pará, Brazil

[7]Coord. Botânica – COBOT, Museu Paraense Emílio Goeldi - MPEG, Campus de Pesquisa, Belém, Pará, Brazil

[8]Lewis B. and Dorothy Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, New York, USA

[9]Department of Biosystems Science and Engineering, ETH Zurich, Zurich, Switzerland

[10]Department of Biological and Environmental Sciences and Gothenburg Global Biodiversity Centre, University of Gothenburg, Gothenburg, Sweden

[11]Department of Biology, University of Oxford, Oxford, UK

[12]Royal Botanic Garden Edinburgh, Edinburgh, UK

Correspondence
Jed Arno, Department of Computer Science, Royal Holloway, University of London, Surrey, TW20 0EX, UK.
Email: jed.arno@rhul.ac.uk

Societal Impact Statement

Understanding and protecting plant life is essential for tackling the twin challenges of biodiversity loss and climate change. To support this, we have developed a new digital approach that helps identify plant species more quickly and accurately. By using images of preserved plant specimens from global collections sourced through the Global Biodiversity Information Facility and combining computer vision technology with expert knowledge from plant scientists, our approach makes it easier to catalogue and study plants. This innovation not only speeds up scientific research but also strengthens the connection between traditional physical plant collections and modern digital collections and tools—helping scientists, conservationists and communities work together to safeguard nature.

Summary

- Computer vision applied to digital herbarium collections holds tremendous promise to streamline specimen identification and accelerate the work of taxonomists and herbarium curators.
- We present a sampling and image preprocessing pipeline applicable to any image dataset that uses the Darwin Core data standard. We tested it on Cyperaceae, a large monocot plant family known for its identification challenges, and on Rhamnaceae, a eudicot plant family, to demonstrate broad applicability across angiosperms.
- Digitised herbarium specimens were sampled via the Global Biodiversity Information Facility to create image datasets with balanced representation annotated with taxon labels. These were used to train deep learning models at genus level in Cyperaceae and Rhamnaceae, and at species level in the genera *Bulbostylis* and *Ziziphus*.

Isabel Larridon, Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, UK.
Email: i.larridon@kew.org

Funding information
Royal Holloway, University of London; Royal Botanic Gardens, Kew; ETH Zurich; Swedish Research Council, Grant/Award Number: 2024-04303; Swedish Foundation for Strategic Environmental Research MISTRA, Grant/Award Number: BIOPATH (F 2022/1448)

- A model fine-tuned on the data performed efficiently and consistently achieved top-1, top-3 and top-5 accuracy rates of ≥72%, ≥88% and ≥92% in identifying digitised herbarium specimens of Cyperaceae and Rhamnaceae to genus level. Species-level identification in *Bulbostylis* reached 65%, 83% and 89%, while *Ziziphus* achieved higher rates of 72%, 85% and 90%. Our approach integrates an automated pipeline for dataset generation with expert verification to enhance data quality. This framework supports scalable, accurate identification of herbarium specimens and fosters a more dynamic relationship between digital and physical collections.

KEYWORDS
computer vision, deep learning, digital herbaria, herbarium specimens, identification, taxonomy

1 | INTRODUCTION

Understanding plant diversity is critical for the development of sustainable nature-based solutions to the biodiversity and climate crises (Antonelli et al., 2020). Accurate identification and delimitation of species, the most widely used component of biodiversity (Coates et al., 2018), is fundamental in this process. Rapid, accurate, scalable and cost-effective species identification and delimitation methods are needed for understudied species-rich plant groups (Grace et al., 2021), particularly in the tropics where most species occur and the threats to plant diversity are most acute. While DNA barcoding fulfils these criteria for some plant groups (Hollingsworth et al., 2016, 2011), a single DNA region does not provide enough evidence for accurate species identification in others (Hollingsworth, 2011), but genome-scale sequencing is not yet cost-effective at the scale of natural history collections nor where laboratory and sequencing resources are scarce. Now, innovations in machine learning enable us to accelerate and democratise species identification, discovery and delimitation in silico, through new tools and workflows using expert-identified herbarium specimens as a reference or training dataset.

The application of deep learning methods—a subset of machine learning that uses large neural networks to learn from data—could provide a major boon to accelerate taxonomic research through the use of convolutional neural networks (CNNs) and vision transformers (ViT) (Arno et al., 2024; Karbstein et al., 2024; Perez et al., 2022). Neural networks are trained to learn and extract useful features from images in the form of image embeddings that can be applied to computer vision tasks such as classification, segmentation, object detection, image restoration and more. These models have shown impressive performance on benchmark tasks (Bello et al., 2021; Dosovitskiy et al., 2021; Liu et al., 2021; Szegedy et al., 2017), demonstrating the quality of the features extracted by these models. To make practical use of advances in deep learning models, high-quality training data for the target are needed, particularly with respect to the quality of the labels; in the herbarium context, correct taxonomic identification of the specimens used as reference for training the algorithm are needed.

There has been extensive research into training computer vision models to classify plant taxa using their leaves and flowers (Apriyanti et al., 2021; Brun et al., 2025; He et al., 2016; Lee et al., 2018; Rzanny et al., 2019, 2022; Seeland & Mäder, 2021). Use of image sets of fresh leaves has been very popular over the last 20 years as benchmark datasets in hundreds of studies (Ahmed et al., 2023). The size and challenge of the available plant datasets has developed from simple datasets such as the Swedish Leaf Dataset (Söderkvist, 2001), which comprises 15 species with 75 images of fresh leaves each, to millions of images of dried specimens representing thousands of species, such as the Half-Earth dataset (de Lutio et al., 2022) (Table 1).

Benchmark datasets have been useful for assessing the performance of computer vision models for classification (i.e., plant identification), while progression in vision models has justified much larger datasets comprising millions of images (Table 1). These have begun to reflect the challenges of automated plant identification, such as the variation in the number of available example specimens between taxa, the prevalence of misidentified specimens and the fine-grained visual differences in challenging taxa that can confound real-world classification tasks.

A simultaneous increase in the availability of digitised herbarium specimens, available under Creative Commons license conditions, and the improvement in the performance of computer vision using deep learning models (de Lutio et al., 2022; Goëau et al., 2022) now allows real-world applications of in silico species identification. Whilst the quest for increasingly accurate, accessible plant identification tools for non-specialists is a continued challenge for the machine learning community (de Lutio et al., 2022; Wäldchen et al.2018), specialist applications are also required for use in herbaria. Millions of herbarium specimen images are becoming available online each year through large digitisation programmes by natural history institutions (e.g., De Smedt et al., 2024; Le Bras et al., 2017) and by national and international projects in which institutions work together such as REFLORA, DiSSCO and iDigBio (Pinheiro et al., 2024; Smith et al., 2022; Thiers, 2024). As a result, the Global Biodiversity Information Facility (GBIF) currently includes images of ~120 million preserved specimens of plants (GBIF.org, 2024a). Three scenarios stand to benefit from

TABLE 1 Development of benchmark datasets generated for training computer vision models for plant identification.

Dataset	Number of images	Categories	Type
Flavia (Wu et al., 2007)	1095	33 species belonging to 17 families	Fresh leaf images for species classification
PlantVillage (Hughes & Salathe, 2016)	54,309	14 crop species from 8 families. There are a mixture of healthy and diseases specimens with a total of 38 categories.	Images of healthy and diseased crop leaves
Leaf-12 (Pearline & Kumar, 2019)	3840	12 species from 12 different families	Fresh leaf images used for species classification
LifeClef 2017 (Goeau et al., 2017) Challenge dataset 1	256,287	10,000 species	Images of living specimens from the Encyclopedia of life (Parr et al., 2014)
LifeClef 2017 Challenge dataset 2	1.1 million	10,000 species	Images of living specimens collected using a web scraper
PlantNet-300k (Garcin et al., 2021)	306,293	1081 species	Crowdsourced images of living specimens
FGVC6 2019 Herbarium Challenge (Tan et al., 2019)	46,000	683 species, all from the Melastomataceae family	Dried specimens from the New York Botanical Garden collection
A benchmark dataset of herbarium specimen images with labelled data (Dillen et al., 2019)	1800	204 families from 58 orders	Dried specimens from 9 institutions
Half-Earth Challenge dataset (de Lutio et al., 2022)	2.5 million	64,500 species from 451 families	Dried specimens from 5 institutions
The Herbarium 2022: Flora of North America NAFlora-1M dataset (Park et al., 2024) for FGVC9 challenge	1.05 million	15,500 North American species	Dried specimens collected from 60 institutions

such innovation: (1) identifying species, (2) finding misidentified specimens in a collection and (3) grouping visually similar specimens to identify or re-evaluate taxon boundaries. Representation learning, in which data classes are automatically identified among large image datasets, can be fine-tuned for set-valued classification to aid identification or to produce similarity scores for automated visual comparisons of specimens.

Here, we present a pipeline for plant identification from digitised herbarium specimens. We validate it at the genus and species ranks in the large monocot family Cyperaceae (>5600 species, 95 genera; Larridon, 2022). Both the family in general and the genus *Bulbostylis* (hair sedges, ~230 species) in particular are taxonomically complex and difficult to identify to species level, due to unclear species limits linked to factors such as wide intraspecific variation and narrow interspecific variation (Larridon et al., 2021; Xanthos et al., 2023). Sedges are key components of open ecosystems such as grasslands and wetlands (e.g., Rasaminirina et al., in press). Besides their ecological importance, sedges are also economically important for a variety of uses (Simpson & Inglis, 2001) or because they are problematic weeds (e.g., Bryson & Carter, 2008). Overall, ~20% of sedge species are Near Threatened or threatened (IUCN, 2025), although in some groups >70% are threatened (e.g., *Costularia*; Larridon et al., 2019). Tools to help identify challenging species-rich plant groups like Cyperaceae will support our understanding, conservation and use of these groups for nature-based solutions to the climate and biodiversity crises. To contrast our results from this challenging monocot group and demonstrate broad applicability across angiosperms, we also tested our pipeline on Rhamnaceae, a eudicot family that is taxonomically well-understood (Richardson, Fay, Cronk, Bowman, & Chase, 2000; Richardson, Fay, Cronk, & Chase, 2000). The genus *Ziziphus* Mill. in this family is a particularly useful test case, as it displays a wide diversity of habit types (climbers, shrubs, and trees), leaf and fruit shapes and indumentum types (Cahen et al., 2021). Our pipeline allows researchers to scale up their ability to develop and test dedicated deep learning models to improve the efficiency and accuracy of taxonomic identification of digitised herbarium specimens.

2 | MATERIALS AND METHODS

2.1 | The pipeline

We developed a pipeline that enables the generation of datasets for training deep learning models for specialised tasks using specimen images and observations provided in the DarwinCore (DwC) format, the data standard for sharing biodiversity information (Wieczorek et al., 2012). Several considerations need to be addressed to effectively use DwC query data as a training set for a deep learning model. (1) Observation records are unbalanced (de Lutio et al., 2022): Images of a plant group will typically comprise a small number of taxa for which many digitised herbarium specimens are available, and many taxa with very few digitised specimens. For example, tens of thousands of images

for a few taxa when the others all have less than 100 would mean that the over-represented taxa have an unnecessary number of images, potentially affecting the performance of the models trained on these data. (2) A practical and useful image dataset should be of a manageable size and provide a more balanced representation of the different taxa. This practice also helps the models to accurately recognise the poorly represented categories. (3) Herbarium specimen repositories contain misidentified specimens and images without plant material, such as capsules and ethnobotanical objects, which may not be filtered out by the query. Removing these images from the final useful image dataset may require automated or manual steps. (4) The model must be prevented from learning features not related to the plant material, so features such as text labels, logos and barcodes need to be blurred. (5) Finally, the images are divided into a training set and dedicated holdout test and validation sets. Since some classes contain very few images, selecting a validation set randomly from the entire set could result in some classes being excluded. In our pipeline, this split is performed class by class to ensure that the validation and test sets maintain the same class distribution as the original set. The holdout test set comprises 10% of the images from each category. Twenty per cent of the remaining images from each class are used for the validation set.

Our pipeline compiles a specimen image dataset from multiple data providers via the GBIF data portal. The number of images per taxon is balanced, text features are blurred and the model is trained to classify the images into taxa (Figure 1). The output of the pipeline consists of two subdirectories, one containing a set of images for model training and a smaller set of images for model evaluation.

2.2 | Data compilation

The first step is to query GBIF.org (2024a) for herbarium specimens. Each query must specify the 'basis of record' as a preserved specimen to select images of herbarium specimens. Desired taxa can be selected and further filtered. For example, 'publisher' and 'location' can be applied if desired. We note that using images from different publishers within the same image dataset introduces further variation as the resolution and layout of specimen images may vary between publishers, that is, herbaria. Differences in image resolution can affect the blurring of the labels, and high-resolution images require more storage, so images were resized to 1090×1600 pixels for consistency.

The DwC contains a lot of information about each observation record. For this pipeline, we are only interested in the unique GBIF ID, the 'identifier' that contains the image link, and the taxonomic labels for the observation record. A table is created with a column for the GBIF ID and the image link. The taxa are separated into 'family', 'genus' and 'species' columns, which allows for an analysis of the distribution of taxa in the observations and a selection of the labels used by the classifier.

2.3 | Sampling from the query data

After selecting the taxonomic level to be used as a label set, a training set can be sampled from the table. When sampling, the aim is to reduce the size of the download and help balance the representation of different taxa. Upper and lower thresholds are defined for the number of images for each taxon. This bounding of the number of images per taxon allows us to (1) prevent the over representation of a small number of taxa, (2) remove taxa with too few images for learning and (3) reduce the range of the number of images for each taxon.

Subsampling is used to enforce the upper threshold. If a taxon has too many images, then a subset of the images belonging to the taxon is randomly selected. If a taxon has fewer images than the lower threshold, it is removed. The resulting sample is thus selected to lead

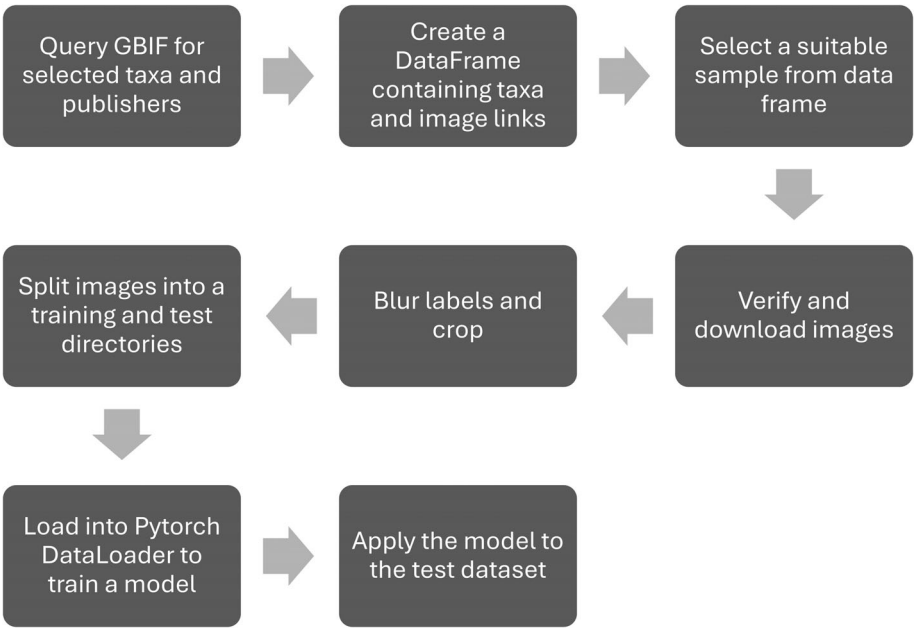

FIGURE 1 Our pipeline for training deep vision models using images from GBIF. All steps are implemented in a user-friendly Python package automating most steps and requiring minimal programming skills.

to high accuracy while maintaining a manageable size to reduce memory requirements and training time. The sampling approach used is given in Algorithm 1.

Algorithm 1: Sample a subset of images

Input : A data frame of image links and labels, D, an upper threshold of samples T_U, a lower threshold for samples T_L

Returns: A sample S where $S \subseteq D$

```
for y ∈ Y do
    // For each unique label
    // If there are more images than the upper threshold
    if frequency(y) > T_U then
        // Then the sample we include a random selection of
           100 images for this category
        S_y = undersample(D_y, T_U);
    // Otherwise, check if there are more images than the
       lower threshold
    else if frequency(y) > T_L then
        // If so then include this category in the sample
        S_y = D_y;
    else
        // If there are less images than the lower threshold,
           exclude this category from the sample
        S_y = ∅;
end for
```

2.4 | Downloading and preparing images

Downloaded images are organised into folders for each label in the training set. Organising images into folders by their label makes it easy to import images when training a model. When downloading images, if there are issues using the image link provided for an observation, it is discarded. If there are unused observations of the same class in the query results, then the observation can be resampled.

Before training a model on the images, some preprocessing needs to be done. A lot of digitised herbarium specimens will contain more visual information than needed. Anything around the herbarium specimen mounting sheet should be cropped out. Any labels or text found on the sheet should be blurred. These steps are crucial to avoid learning spurious features and ensure the model makes predictions using the features of the plant material.

Blurring the labels found on specimens is performed in three stages: (1) using the CRAFT text detection model (Baek et al., 2019) to identify text; (2) collecting and bounding clusters of text; and (3) applying a Gaussian Blur to the boundaries. CRAFT stands for Character-Region Awareness For Text Detection. Using VGG18 (Simonyan & Zisserman, 2014) as a backbone, features are extracted at several points to identify region scores and affinity scores. The region score is used for identifying characters, and the affinity score is used to group characters into regions of text. Using CRAFT to construct bounding boxes around text identifies regions at a 'word level'. These initial bounding areas may leave gaps on the text labels. Detected text regions that are close together are grouped and merged so that areas of the image that contain text can be blurred.

Lastly, images contain additional annotations around the specimen that would be useful to a human observer. Since we are interested in training a model on visual features found on the plant specimen, we crop around the herbarium specimen mounting sheet to remove any additional objects or background found in the image. The blurring process has been tested on specimens belonging to different families and from different publishers to show that the blurring process works consistently. The main consideration is the resolution of the image, which will affect the best margin value to use, as the margin is relative to the number of pixels.

2.5 | Expert verification

GBIF is the definitive aggregator for digital natural history collections, and the vast data are sourced from a broad range of publishers (i.e., herbaria). Observations downloaded may contain errors and should be checked by an expert before training a model. This is the main step in our pipeline that is not automated but requires human input. Expert verification was carried out by (1) visually inspecting all large size thumbnail images for any issues identified in the thumbnail images checked at full resolution; and additionally (2) checking 10 random images in detail at full resolution per genus (Cyperaceae and Rhamnaceae datasets) or species (*Bulbostylis* and *Ziziphus* datasets).

2.6 | Application

To test the value of the pipeline presented here, the Cyperaceae image dataset was used to train three classifiers using three ViT architectures: ViT-b/16, ViT-l/16 and ViT-h/14 (Table 2). The initial pretrained weights are from semi-supervised training (Cai et al., 2022) on ImageNet1k (Deng et al., 2009). An input resolution of $224 \times 224 \times 3$ is used for all models to match the input resolution used for pre-training.

To deal with any remaining imbalance in the dataset, a weighted sampling method was used in PyTorch (Paszke et al., 2019). The probability of selecting an image during training is adjusted by how well

TABLE 2 ViT architecture settings used for specimen classification.

Model	Number of tokens (including class tokens)	Number of hidden layers	Number of encoder blocks	Number of parameters	Estimated Total size (MB)
ViT-b/16	197	768	12	86,567,656	3582.53
ViT-l/16	197	1024	24	304,326,632	9611.91
ViT-h/14	257	1280	32	632,045,800	20738.85

each species is represented. The weights for each image are given in the equation. The weights w_i of a sample x_i belonging to a class C is the total number of samples in the dataset N divided by the number of samples in the class N_c.

$$w_i = \frac{N}{N_c} \quad (1)$$

Model training was performed on an NVIDIA L4 Tensor Core GPU using Google Colab. The AdamW optimiser (Loshchilov & Hutter, 2019) was used with a learning rate of 0.001, beta values of 0.95 and 0.999 and a weight decay of 0.01. Label smoothing cross entropy (Szegedy et al., 2016) was used as the loss function, owing to the fact that it is widely adopted in existing studies (Müller et al., 2019). The classification head and the final encoder block were trained using 12 epochs. The performance of the models was evaluated using the top-1, top-3 and top-5 accuracy on the holdout test data. Top-*k* accuracy measures how often the correct taxon name is retrieved in the *k* most likely taxa predicted by the model. Consequently, top-1 accuracy requires the model's single most likely prediction to be correct, while top-3 accuracy considers the correct answer to be correct if it is in the model's top three predictions, and top-5 accuracy reflects if the correct answer is among the top 5.

3 | RESULTS

3.1 | The pipeline

The pipeline is implemented as a collection of tools for sampling images from a DwC that can be applied to any GBIF query. The code is available at https://github.com/jedarno/DarwinCoreToDataSetTools. The tools include taxon frequency analysis at different taxonomic levels, sampling of subsets, blurring of labels and preparation of training and holdout test data. This pipeline is used to produce a dataset for training a vision transformer for specimen identification purposes.

3.2 | Data compilation

Four image datasets were compiled using the pipeline: genus-level datasets for Cyperaceae (GBIF.org, 2024b) and Rhamnaceae (GBIF.org, 2025a) from the Kew Herbarium, and two species-level datasets, one of the Cyperaceae genus *Bulbostylis* (GBIF.org, 2024c) from all publishers (herbaria) and of the Rhamnaceae genus *Ziziphus* (GBIF.org, 2025b) from the Kew Herbarium. The data used for image datasets were collected using GBIF queries. The queries contained image links for the specimens and information about the observation, including a taxonomic description broken down into separate columns for each rank. The desired data frame contained one column of image links and another column for the desired labels. For the Cyperaceae and Rhamnaceae images, the label column was the genus, and for the *Bulbostylis* and *Ziziphus* image datasets, the label was the species.

TABLE 3 GBIF query for preserved Cyperaceae and Rhamnaceae specimens published by the Royal Botanical Gardens, Kew.

GBIF field	Family 1	Family 2
Scientific name	Cyperaceae Family	Rhamnaceae Family
Basis of record	Preserved specimen	Preserved specimen
Publisher	Royal Botanic Gardens, Kew	Royal Botanic Gardens, Kew

TABLE 4 GBIF query for preserved *Bulbostylis* from any source publisher and *Ziziphus* specimens published by the Royal Botanical Gardens, Kew.

GBIF field	Cyperaceae genus	Rhamnaceae genus
Scientific name	*Bulbostylis* Kunth	*Ziziphus* Mill.
Basis of record	Preserved specimen	Preserved specimen
Publisher	-	Royal Botanic Gardens, Kew

The filters used for collecting images from GBIF are provided in Tables 3 and 4. To collect images for the Cyperaceae family from the Kew Herbarium, the filters in Table 3 are used. This query returns all observations of preserved specimens from the Kew Herbarium that belong to the Cyperaceae family. To collect images for the genus *Bulbostylis* from all publishers, that is, all herbaria, the filters in Table 4 are used. This query returns all observations of preserved specimens belonging to the genus *Bulbostylis* of the Cyperaceae family.

Similar queries were used to collect the Rhamnaceae images. Images from the Kew Herbarium were used to create a Rhamnaceae dataset to be classified to genus level, and a dataset of the genus *Ziziphus* to be classified by species. Unlike with *Bulbostylis*, there were enough *Ziziphus* images in the Kew Herbarium collection to create a dataset, so adding images from other publishers was unnecessary.

The Cyperaceae query (Table 3) returned 156,121 observations. Not all observations were identified to genus level, and some were missing image links. After creating a data frame with the image links and their taxonomic description, all entries with null values were removed using the Pandas method dropna, leaving 145,662 records. The *Bulbostylis* query (Table 4) returned 70,075 observations. After creating a dataframe and removing all records with any null values, 39,694 records remained. The Rhamnaceae query (Table 3) returned 20,249 observations, and the *Ziziphus* query (Table 4) returned 3345 observations from the Kew Herbarium. For both Rhamnaceae queries, no records needed to be removed due to null values.

3.3 | Sampling from the query data

A subset of the Cyperaceae and *Bulbostylis* image datasets was selected using Algorithm 1 with an upper threshold of 100 images and a lower threshold of 20. The selected Cyperaceae sample contains 5092 images belonging to 65 genera with an average of 79.6 images

per genus and a standard deviation of 28.4. Twenty-seven genera represented by fewer than 20 images were removed. At the genus level, the *Bulbostylis* query initially returned observations of 195 species, but many species did not meet the lower threshold of 20 images and were removed. The final subset of *Bulbostylis* images contained 5954 images belonging to 87 species, with an average of 68.4 images per species and a standard deviation of 32.7.

Some of the image links provided did not work. The image links were verified automatically and where possible broken image links were replaced with another image from the same category. There were 11 cases in the Cyperaceae image dataset where broken links could not be replaced leaving 5081 images in the Cyperaceae image dataset. The initial *Bulbostylis* download from all publishers included 105 broken image links that could not be replaced, leaving a total of 5849 images.

Initially, the distribution of images-per-genus in the Cyperaceae image dataset was very imbalanced (Figure 2a). Some genera had only one example image whilst others had tens of thousands. For optimal model training, it is important that each category has sufficient examples but having too many images in some categories can introduce bias and create unnecessary storage requirements. This was resolved after applying the sampling method from Algorithm 1 to sample the Cyperaceae images (Figure 2b). The resulting data frame contained image links and taxonomic labels suitable for model training, and a weighted random sampling method was applied during training to address the remaining imbalance in images-per-genus. The distribution of species in the *Bulbostylis* images available on GBIF shows a similar distribution to the Cyperaceae images, with a longer tail, which was similarly addressed by sampling the distribution of images (Table 3, Figure 2c,d).

The same threshold values were used for sampling a subset of the Rhamnaceae images. The Rhamnaceae sample contains 3751 images belonging to 45 genera with an average of 83.4 images per genus and a standard deviation of 26.7 after removing 18 genera with less than 20 example images (Figure 2e,f). The *Ziziphus* image dataset contained 1654 images belonging to 27 species with an average of 61.3 images and a standard deviation of 31.3 (Figure 2g,h). Thirty species were removed due to having less than 20 images.

3.4 | Preparing downloaded images

Label blurring and cropping were applied to all downloaded images for Cyperaceae and *Bulbostylis*. The barcode and labels were blurred, and anything around the mounting sheet was cropped out (Figure 3).

3.5 | Expert verification

The accuracy of digital specimen identification varies due to the variation in the preparation of herbarium specimens, taxonomic progress, and the many different sources of records (with varying quality control and data quality standards) in aggregators such as GBIF. As such, images downloaded from GBIF may be misidentified, resulting in incorrectly labelled images. Although misidentifications are challenging to spot by eye among potentially large image sets, particularly at species rank, expert verification reduces the risk of perpetuating inaccuracies if a model is trained on inaccurate data. How damaging the incorrect labels will be to the analyses based on the image dataset depends on the size of the training set and the visual diversity of the taxon.

Issues relating to the provided image link, such as corrupted files or duplicate files, were identified when selecting a sample from the downloaded query. Each link is verified automatically before downloading. Any invalid links were then discarded and replaced with a different observation from the GBIF query. Post-analysis, we also tested automation of this process (Notes S1).

Expert verification of the Cyperaceae image dataset of 5081 images identified 42 problematic images, which were manually deleted after the image dataset was downloaded, resulting in a final Cyperaceae image dataset of 5039 images (Table S1). Issues encountered included misidentifications, images representing ethnobotanical objects, photographic vouchers, insufficient plant material or non-representative features (e.g., roots only) and the plant material being hidden within a capsule on the herbarium sheet. Of the 5849 *Bulbostylis* images, 913 were removed, providing a verified dataset of 4936 images (Table S2). Part of the reason for the higher number of problematic images in this dataset was the presence of duplicate images, an issue that did not occur in the Cyperaceae image dataset, which was restricted to specimens from the Kew Herbarium. Including herbarium specimen images from all publishers (herbaria) provided a much larger image dataset but also resulted in a higher percentage of problematic images. For the Rhamnaceae images at genus (Table S3) and species level (Table S4), no such intervention was required. No images were identified for removal.

3.6 | Application

After verifying the quality of the datasets, we used them to train three ViT architectures for classification tasks. The three ViT architectures were trained and tested on our Cyperaceae and Rhamnaceae datasets with 10% of the images from each category set aside for testing. For both families, the models were trained for two tasks. The first task involved classifying images from each family into their respective genera. The second task used images from a specific genus (*Bulbostylis* and *Ziziphus*) and classified them at the species level. The performance of the ViT architectures is compared to the CNN architecture ResNet50 (He et al., 2016).

The models were trained for classifying the Cyperaceae images into 65 genera and classifying Cyperaceae *Bulbostylis* images into 87 different species. The performance of each model on classifying the Cyperaceae specimens at genus level, and of classifying *Bulbostylis* specimens at the species level is provided in Table 5.

Additionally, the models were trained for classifying the Rhamnaceae images into 55 genera and classifying *Ziziphus* images into

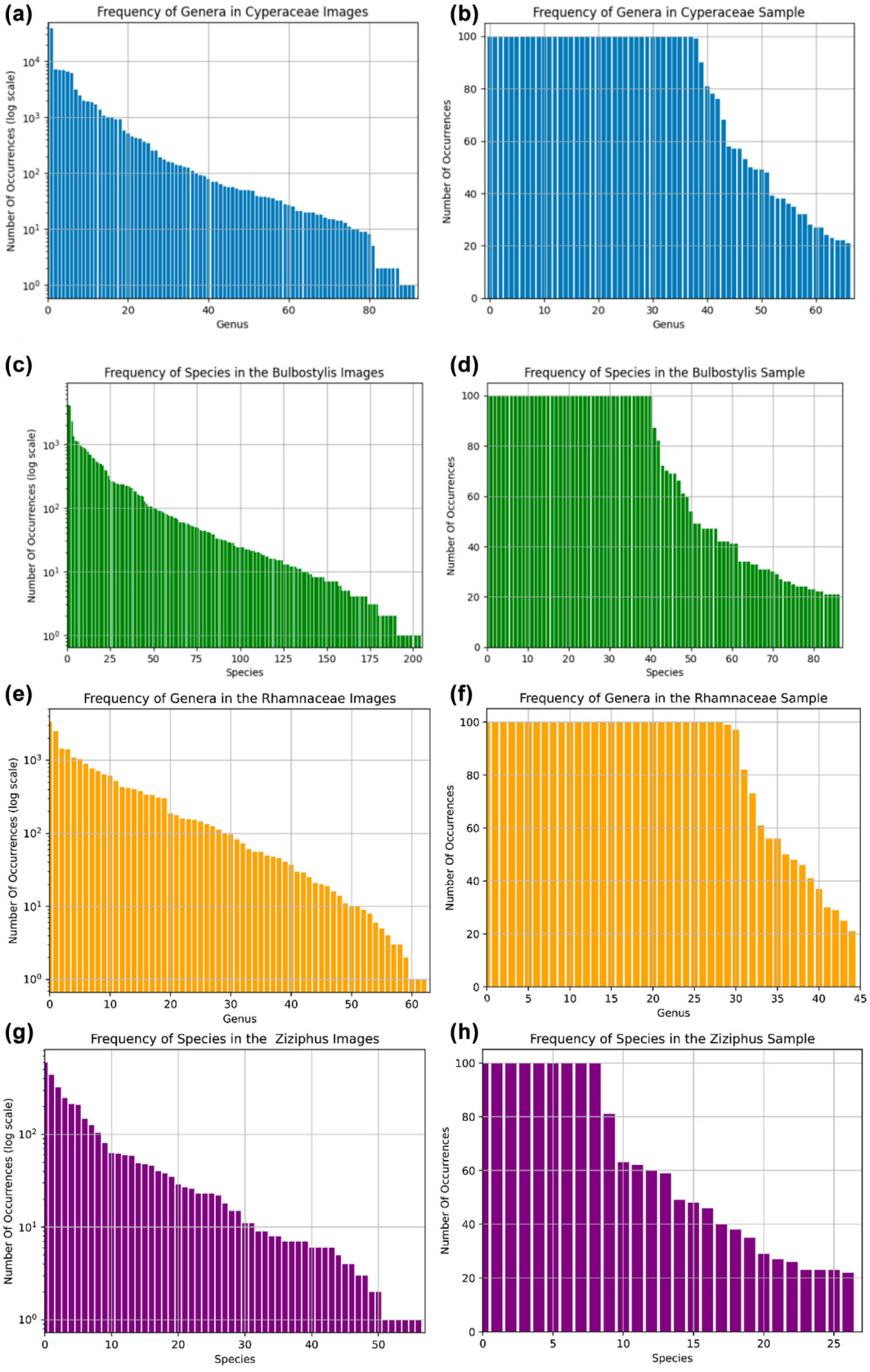

FIGURE 2 Frequency of images for the genera or species in the image datasets in descending order. (a) Cyperaceae initial query, (b) Cyperaceae sampled subset, (c) *Bulbostylis* initial query, (d) *Bulbostylis* sampled subset, (e) Rhamnaceae initial query, (f) Rhamnaceae sampled subset, (g) *Ziziphus* initial query and (h) *Ziziphus* sampled subset. Note that the y-axis is log-scaled in (a, c, e, g).

FIGURE 3 *Bulbostylis* specimen image sourced from GBIF from the Kew Herbarium (K) (a) before and (b) after blurring was applied.

TABLE 5 Classification performance of the models on the holdout test images of Cyperaceae representing 65 genera, and of *Bulbostylis* representing 87 species.

Dataset	Model	Top-1 accuracy	Top-3 accuracy	Top-5 accuracy
Cyperaceae	ResNet50	0.4906	0.6981	0.7862
	ViT-B/16	0.7023	0.8616	0.9057
	ViT-L/16	0.7505	0.8784	0.9182
	ViT-H/14	0.7421	0.8868	0.9224
Bulbostylis	ResNet50	0.3978	0.6516	0.7355
	ViT-B/16	0.5849	0.8129	0.8882
	ViT-L/16	0.6528	0.8344	0.8860
	ViT-H/14	0.6538	0.8258	0.8796

27 different species. The performance of each model on classifying the Rhamnaceae specimens at genus level and of classifying *Ziziphus* specimens at the species level is provided in Table 6. The models performed slightly better on the Rhamnaceae dataset than the Cyperaceae dataset (Table 7).

A confusion matrix illustrates the prediction results for each Cyperaceae genus using the most performant model, that is, ViT-L/16 (Table 5), with the correct predictions shown along the diagonal (where the predicted identifications equal the true identifications). Figure 4a demonstrates that most Cyperaceae genera are accurately identified, with high diagonal values. Misidentifications can be seen between genera that are visually similar, for example, *Schoenus* L. and *Tetraria* P.Beauv. These two genera have recently been redelimited based on molecular studies (Larridon et al., 2018) and are difficult to distinguish based on morphology alone. A second confusion matrix shows prediction results for the Rhamnaceae genera using the most performant model, that is, ViT-L/16 (Table 6), further confirming its effectiveness (Figure 4b). However, a few Rhamnaceae genera are shown to be particularly challenging to identify, that is, *Ampelozizyphus* Ducke, *Noltea* Rchb. and *Retanilla* (DC.) Brongn. These three genera are only represented by 17–22 training images resulting in very few holdout test examples.

TABLE 6 Classification performance of the models on the holdout test images of Rhamnaceae representing 55 genera, and of *Ziziphus* representing 27 species.

Dataset	Model	Top-1 accuracy	Top-3 accuracy	Top-5 accuracy
Rhamnaceae	ResNet50	0.5811	0.7757	0.8486
	ViT-B/16	0.6811	0.8811	0.9486
	ViT-L/16	0.7216	0.8946	0.9297
	ViT-H/14	0.7216	0.8784	0.9135
Ziziphus	ResNet50	0.6624	0.8535	0.8917
	ViT-B/16	0.7261	0.8471	0.8981
	ViT-L/16	0.7006	0.8599	0.9045
	ViT-H/14	0.6943	0.8662	0.9045

TABLE 7 The performance of ViT-L/16 across the different frequency values for the different genera in the Cyperaceae holdout test images and the Rhamnaceae holdout test images.

Dataset	Frequency group	Average accuracy	Average precision	Average recall
Cyperaceae	Low frequency (20–49 training images images)	0.9408	1.0000	0.9254
	Medium frequency (50–89 images)	0.7247	0.9698	0.7241
	High frequency (90–100 images)	0.7228	0.7870	0.7216
Rhamnaceae	Low frequency (20–49 training images images)	0.7073	0.9268	0.7073
	Medium frequency (50–89 images)	0.8095	1.0000	0.8095
	High frequency (90–100 images)	0.7175	0.7594	0.7175

Additional training samples are likely needed to improve performance for these cases.

To assess performance across genera with varying representation in the training data, categories were grouped into three frequency levels based on image count: low (20–49), medium (50–89) and high (90–100). For each group, the average accuracy, precision and recall of ViT-L/16 on holdout test images are reported in Table 7. Precision, defined as the ratio of true positives to total positive predictions, reflects the accuracy of positive predictions, while recall measures the model's ability to identify all positive samples. Surprisingly, the low-frequency group performed best in the Cyperaceae dataset, with exceptionally high precision and recall—indicating no false positives. This may be due to reduced variation in the training images. For Rhamnaceae, the model showed slightly lower performance for low and high-frequency groups compared to medium. This may stem from redundant details in high-frequency classes and limited cues in low-frequency ones.

4 | DISCUSSION

4.1 | A pipeline for automated plant identification

Our pipeline takes advantage of the rapidly increasing number of digital herbarium specimens in public repositories like GBIF to generate datasets for training deep learning models for plant specimen identification. Attributes incorporated into the pipeline such as blurring labels and text annotations, and balancing representation of images per taxon contributed to top-5 identification accuracy rates of up to 0.9224 (92%) at genus (Table 5) and 0.8860 (89%) at species level (Table 6), respectively. This rate compares favourably to other CNN-based approaches that have achieved accuracy rates exceeding 85% in recent years (Arno et al., 2024; Carranza-Rojas et al., 2017; Shirai et al., 2022). This is owing to the efficiency of the pre-processing steps (e.g., text removal and blurring) and superior feature learning capabilities of the selected fine-tuned transformer models.

The pipeline can be rapidly run when connected to the internet and with access to a GPU, or services such as Google Colaboratory. The pipeline is applicable to any plant group for which there are herbarium specimen images on GBIF or other repositories using the Darwin Core data standard. Methods are provided in the pipeline that can be used to sample manageable datasets from large query results, remove supplementary material such as labels and barcodes and provide an easy-to-use directory of images for model training. These directories can be easily used with Python libraries such as PyTorch (Paszke et al., 2019) and Keras (Chollet, 2015) that, in turn, make deep vision models easily accessible in the specimen identification process.

The extraordinary rate at which deep learning models have developed for extracting useful image features now permits their application beyond vision tasks. Deep computer vision models such as ViT (Dosovitskiy et al., 2021) and SWIN (Liu et al., 2021) produce versatile image representations that can be trained for set-valued classification for identification or similarity scores for visual comparisons such as those used in face recognition problems (Deng et al., 2019; Schroff et al., 2015). Trained to learn useful representations in more specific contexts, such as recognising similarity among digital herbarium specimens, such models are highly effective.

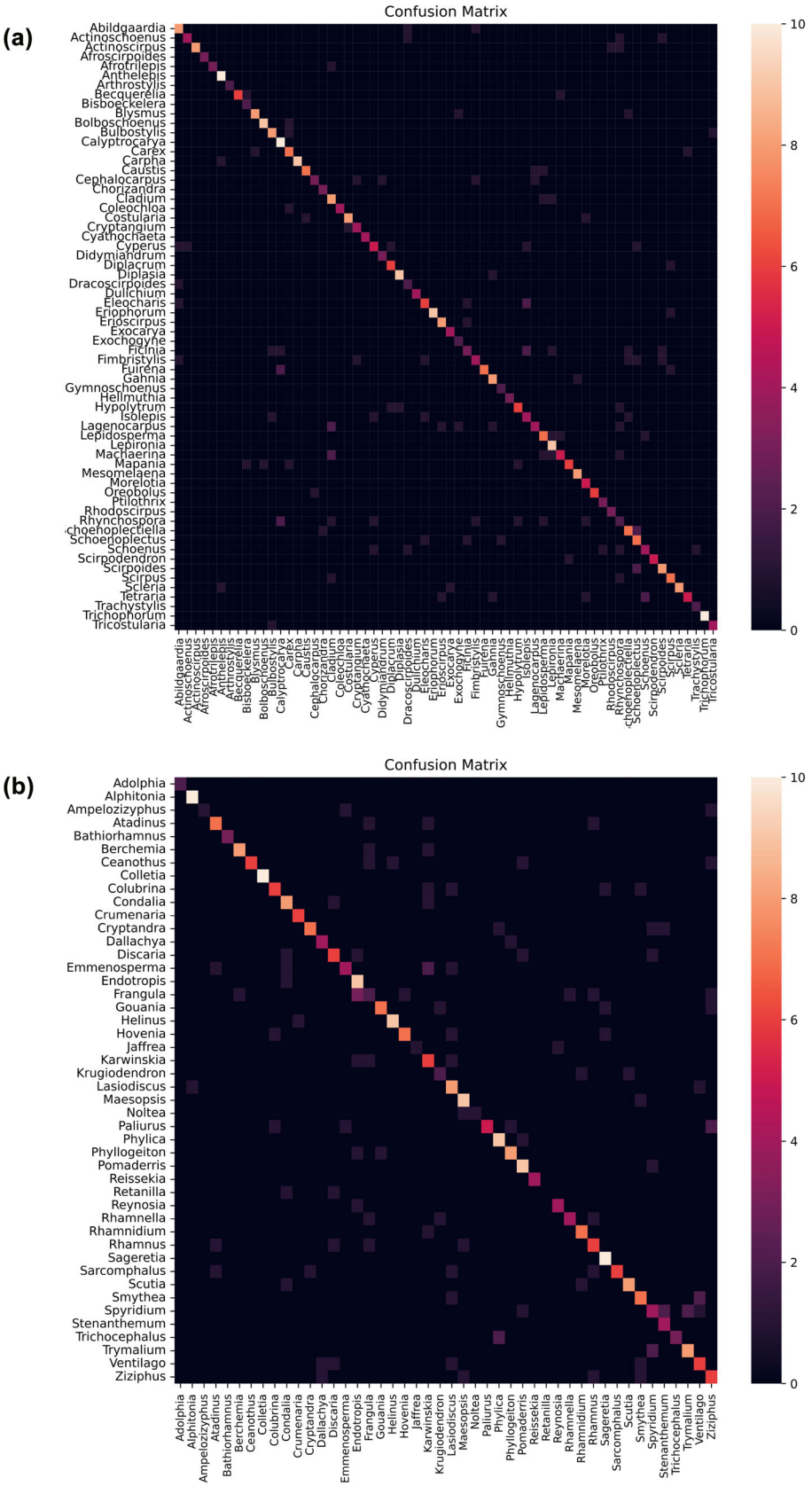

FIGURE 4 (a) Confusion matrix of the ViT-L/16 predictions on the genera of the Cyperaceae holdout test images and (b) confusion matrix of the ViT-L/16 predictions on the genera of the Rhamnaceae holdout test images.

Despite the large and increasing number of digitised herbarium specimens publicly available online and the progress in the development of deep learning models, barriers exist for those who lack specialist analyst skills or have no access to the infrastructure needed to use these resources (Groom et al., 2023). Our pipeline contributes to democratising access to machine learning for specimen

identification and addresses barriers facing plant taxonomists and herbarium curators who wish to adopt machine learning approaches to expedite their research or collections management.

While much of the pipeline is automated, the manual expert verification step is critical to ensure the quality and accuracy of the image datasets. This is even more critical in case of taxonomically complex plant families which are hard to identify such as Cyperaceae. Previous studies reported that plant groups like Cyperaceae, Poaceae or bryophytes, of which images often lack the distinguishing characters needed for fine-scale identification, can have a significant effect on the accuracy of species identification (White et al., 2023). Expert verification at genus level in the Cyperaceae dataset identified <1% problematic images, while expert verification at species level in the *Bulbostylis* image dataset identified 15.6% problematic images (Notes S1, Table S1) highlighting the importance of this step, particularly at the species level and when including images from across multiple publishers (herbaria) on GBIF. Our results show that expert-verified training image datasets, even of a very complex taxonomic group at species level, can achieve a favourable accuracy rate for identification.

4.2 | Application scenarios to accelerate taxonomy and enhance herbarium curation

Making reliably accurate identification models available for specialists to enhance and accelerate their work will help tackle the taxonomic impediment and have a net positive effect on understanding plant diversity (Engel et al., 2021; Gorneau et al., 2022). Scenarios for use of these tools allow the deployment of deep vision methods to specific target problems, be it identification of specimens of a challenging taxon, specimens from a geographical region or the curation of specimens to support naming and determination. Our empirical test on Cyperaceae and Rhamnaceae at genus level and on *Bulbostylis* and *Ziziphus* at species level showed that the automated specimen localisation process along with background noise removal serves as an effective pre-processing technique for specimen image quality enhancement for most application scenarios. The pipeline can handle background text and other noise removal for specimen images with various lighting conditions, image resolutions and background contrast.

Potential application scenarios for taxonomic research and specimen curation:

1. Identification of plant specimens to genus or species level: Using an expert-verified image dataset, the pipeline can be applied to unidentified herbarium specimens to provide an ordered list of possible identifications starting with the most likely identification. For taxa already represented by substantial numbers of digitised herbarium specimens, this 'possible identifications' list can accelerate specimen identification by the expert. Our preliminary experimental example, applying the pipeline for the identification of Cyperaceae and Rhamaceae at genus level, and in *Bulbostylis* and *Ziziphus* at species level using the processed dataset, showed promising performances. However, there is room for further enhancement: the algorithms could benefit from the incorporation of additional metadata (e.g., distribution, phenology, traits, etc.) to further increase accuracy rates.
2. Locating misidentified specimens: For specimens in existing collections, the current identification can be used as an indicator of potentially misidentifications against the model's output. If the model considers the current identification unlikely, specimens can be quickly flagged for re-identification.
3. Finding taxa that are putatively new to science: The feature encodings learned by these models can be used to automatically compare visual similarity among specimens using similarity metrics or clustering methods. This step is automatable when visual comparison is undertaken on specimen image embeddings. Specimens identified as dissimilar to the others in their category could help detect misidentified specimens, support species delimitation, and even aid in discovering species new to science.

4.3 | Future prospects and challenges

On-going digitisation efforts, evolving methods, and the dynamic nature of taxonomy jointly suggest that the relationship between physical and digital herbarium collections is evolving: the physical specimen is no longer the de facto 'accurate' reference but rather can be refined through analysis of the digitised herbarium specimen. In this context, the most urgent challenges confronting the herbarium curator and collections organisations now include resourcing the digitisation of herbaria, updating identification workflows to include deep learning approaches and providing training to curators and collection users to address skills gaps. Herbarium curation can now include deep learning steps to locate specimens requiring an updated determination or label. The taxonomic process will be accelerated and enhanced with human supervisors assisted by deep learning approaches such as this. The role of the specialist taxonomist remains, as ever, critical for manifold tasks, including assessing accuracy; but it should now also include training deep learning models to accelerate species identifications.

AUTHOR CONTRIBUTIONS

Jed Arno, Olwen Grace, Li Zhang and Isabel Larridon conceived and designed the research. Jérémie Morel, Fitiavana Rasaminirina, Juliene de Fátima Maciel-Silva, Daniel Cahen and Isabel Larridon carried out the expert verification of the generated image datasets. Jed Arno designed the pipeline, processed the data, performed the analyses and wrote the manuscript. Li Zhang supervised the pipeline design, analyses and the manuscript preparation. Daniel Cahen, Daniele Silvestro, Alexandre Antonelli, Olwen Grace and Isabel Larridon contributed to the manuscript writing. All authors contributed to the interpretation of results and formulation of conclusions.

ACKNOWLEDGEMENTS

This research is part of the PhD of the first author funded through a College Industrial Match Funded Studentship Scheme between Royal

Holloway, University of London, and the Royal Botanic Gardens, Kew. D.S. received funding from ETH Zurich. D.S. and A.A. also acknowledge funding from the Swedish Research Council (VR: 2024-04303 and 2024-04303, respectively) and the Swedish Foundation for Strategic Environmental Research MISTRA within the framework of the research programme BIOPATH (F 2022/1448). A.A. acknowledges further financial support from RBG Kew Development.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

The code for the pipeline is available at https://github.com/jedarno/DarwinCoreToDataSetTools. The pipeline was tested on digitised herbarium specimens sourced from GBIF (Cyperaceae: https://doi.org/10.15468/dl.7qype7; *Bulbostylis*: https://doi.org/10.15468/dl.q95pvs; Rhamnaceae: https://doi.org/10.15468/dl.zkqd4w; *Ziziphus*: https://doi.org/10.15468/dl.pwu45u).

ORCID

Jed Arno https://orcid.org/0009-0008-9810-9907
Jérémie Morel https://orcid.org/0000-0001-7969-2609
Fitiavana Rasaminirina https://orcid.org/0000-0003-0162-7975
Juliene de Fátima Maciel-Silva https://orcid.org/0000-0002-5725-4170
Daniel Cahen https://orcid.org/0000-0003-4549-7092
Damon P. Little https://orcid.org/0000-0001-9635-6164
Daniele Silvestro https://orcid.org/0000-0003-0100-0961
Alexandre Antonelli https://orcid.org/0000-0003-1842-9297
Olwen Grace https://orcid.org/0000-0003-1431-2761
Li Zhang https://orcid.org/0000-0001-6674-692X
Isabel Larridon https://orcid.org/0000-0003-0285-722X

REFERENCES

Ahmed, S. U., Shuja, J., & Tahir, M. A. (2023). Leaf classification on Flavia dataset: A detailed review. *Sustainable Computing Informatics & Systems*, *40*, 100907. https://doi.org/10.1016/j.suscom.2023.100907

Antonelli, A., Hiscock, S., Lennon, S., Simmonds, S., Smith, R. J., & Young, B. (2020). Protecting and sustainably using the world's plants and fungi. *Plants, People, Planet*, *2*, 368–370. https://doi.org/10.1002/ppp3.10150

Apriyanti, D. H., Spreeuwers, L. J., Lucas, P. J. F., & Veldhuis, R. N. J. (2021). Automated color detection in orchids using color labels and deep learning. *PLoS ONE*, *16*, e0259036. https://doi.org/10.1371/journal.pone.0259036

Arno, J., Grace, O., Larridon, I., & Zhang, L. (2024). Plant species classification using evolving ensemble and Siamese networks. 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 4908–4915. https://doi.org/10.1109/SMC54092.2024.10831570

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, 9357–9366. https://doi.org/10.1109/CVPR.2019.00959

Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., Shlens, J., & Zoph, B. (2021). Revisiting ResNets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, *34*, 22614–22627.

Brun, P., de Witte, L., Popp, M. R., Zurell, D., Karger, D. N., Descombes, P., de Lutio, R., Wegner, J. D., Bornand, C., Eggenberg, S., Olevski, T., & Zimmermann, N. E. (2025). Florid—A nationwide identification service for plants from photos and habitat information. *Environmental Modelling & Software*, *188*, 106402. https://doi.org/10.2139/ssrn.4830448

Bryson, C. T., & Carter, R. (2008). The significance of Cyperaceae as weeds. In R. F. C. Naczi & B. A. Ford (Eds.), *Sedges: Uses, diversity and systematics of the Cyperaceae. Monographs in systematic botany from the Missouri Botanical Garden* (Vol. 108) (pp. 15–101). Missouri Botanical Garden.

Cahen, D., Rickenback, J., & Utteridge, T. M. A. (2021). A revision of *Ziziphus* (Rhamnaceae) in Borneo. *Kew Bulletin*, *76*, 767–804. https://doi.org/10.1007/s12225-021-09970-3

Cai, Z., Ravichandran, A., Favaro, P., Wang, M., Modolo, D., Bhotika, R., Tu, Z., & Soatto, S. (2022). Semi-supervised vision transformers at scale. *Advances in Neural Information Processing Systems*, *35*, 25697–25710. https://papers.nips.cc/paper_files/paper/2022/file/a4a1ee071ce0fe63b83bce507c9dc4d7-Paper-Conference.pdf

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*, 181. https://doi.org/10.1186/s12862-017-1014-z

Chollet, F. (2015). Keras GitHub repository. September 15, 2024. from https://github.com/fchollet/keras

Coates, D. J., Byrne, M., & Moritz, C. (2018). Genetic diversity and conservation units: Dealing with the species-population continuum in the age of genomics. *Frontiers in Ecology and Evolution*, *6*, 165. https://doi.org/10.3389/fevo.2018.00165

de Lutio, R., Park, J. Y., Watson, K. A., D'Aronco, S., Wegner, J. D., Wieringa, J. J., Tulig, M., Pyle, R. L., Gallaher, T. J., Brown, G., Guymer, G., Franks, A., Ranatunga, D., Baba, Y., Belongie, S. J., Michelangeli, F. A., Ambrose, B. A., & Little, D. P. (2022). The herbarium 2021 half-Earth challenge dataset and machine learning competition. *Frontiers in Plant Science*, *12*, 787127. https://doi.org/10.3389/fpls.2021.787127

De Smedt, S., Bogaerts, A., De Meeter, N., Dillen, M., Engledow, H., Van Wambeke, P., Leliaert, F., & Groom, Q. (2024). Ten lessons learned from the mass digitisation of a herbarium collection. *PhytoKeys*, *244*, 23–37. https://doi.org/10.3897/phytokeys.244.120112

Deng, J., Dong, W., Socher, R., Li, L-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 5962–5979. https://doi.org/10.1109/TPAMI.2021.3087709

Dillen, M., Groom, Q., Chagnoux, S., Güntsch, A., Hardisty, A., Haston, E., Livermore, L., Runnel, V., Schulman, L., Willemse, L., Wu, Z., & Phillips, S. (2019). A benchmark dataset of herbarium specimen images with label data. *Biodiversity Data Journal*, *7*, e31817. https://doi.org/10.3897/BDJ.7.e31817

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. Retrieved October 10, 2024 from, [WWW document] https://openreview.net/forum?id=YicbFdNTTy

Engel, M. S., Ceríaco, L. M. P., Daniel, G. M., Dellapé, P. M., Löbl, I., Marinov, M., Reis, R. E., Young, M. T., Dubois, A., Agarwal, I., Lehmann, A. P., Alvarado, M., Alvarez, N., Andreone, F., Araujo-Vieira, K., Ascher, J. S., Baêta, D., Baldo, D., Bandeira, S. A., … Zacharie, C. K. (2021). The taxonomic impediment: A shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of*

the Linnean Society, *193*, 381–387. https://doi.org/10.1093/zoolinnean/zlab072

Garcin, C., Joly, A., Bonnet, P., Lombardo, J-C., Affouard, A., Chouet, M., Servajean, M., Lorieul, T., & Salmon, J. (2021). Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution. NeurIPS 2021-35th Conference on Neural Information Processing Systems, Dec 2021, Virtual Conference, France. https://doi.org/10.5281/zenodo.5645731hal-03474556v2

GBIF.org. (2024a). GBIF home page. Retrieved September 13, 2024, from https://www.gbif.org

GBIF.org. (2024b). [Cyperaceae]. GBIF Occurrence Download. Retrieved June 06, 2024, from https://doi.org/10.15468/dl.7qype7

GBIF.org. (2024c). [Bulbostylis]. GBIF Occurrence Download. https://doi.org/10.15468/dl.q95pvs [accessed 10 June 2024].

GBIF.org. (2025a). [Rhamnaceae]. GBIF Occurrence Download. https://doi.org/10.15468/dl.zkqd4w [accessed 09 September 2025].

GBIF.org. (2025b). [Ziziphus]. GBIF Occurrence Download. https://doi.org/10.15468/dl.pwu45u [accessed 09 September 2025].

Goëau, H., Bonnet, P., & Joly, A. (2022). Overview of PlantCLEF 2022: Image-based plant identification at global scale. CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3180/paper-153.pdf

Goeau, H., Bonnet, P., & Joly, A. (2017). Plant identification based on noisy web data: The amazing performance of deep learning (LifeCLEF 2017). CLEF: Conference and Labs of the Evaluation Forum, Sep 2017, Dublin, Ireland. hal-01629183 https://hal.science/hal-01629183v1/document

Gorneau, J., Ausich, W., Bertolino, S., Bik, H., Daly, M., Demissew, S., Donoso, D., Folk, R., Freire-Fierro, A., Ghazanfar, S., Grace, O., Hu, A., Kulkarni, S., Lichter-Marck, I., Lohmann, L., Malumbres-Olarte, J., Muasya, A., Pérez González, A., Singh, Y., … Esposito, L. (2022). Framing the future for taxonomic monography: Improving recognition, support, and access. *Bulletin of the Society of Systematic Biologists*, *1*(1). https://doi.org/10.18061/bssb.v1i1.8328

Grace, O. M., Pérez-Escobar, O. A., Lucas, E. J., Vorontsova, M. S., Lewis, G. P., Walker, B. E., Lohmann, L. G., Knapp, S., Wilkie, P., Sarkinen, T., Darbyshire, I., Lughadha, E. N., Monro, A., Woudstra, Y., Demissew, S., Muasya, A. M., Díaz, S., Baker, W. J., & Antonelli, A. (2021). Botanical monography in the anthropocene. *Trends in Plant Science*, *26*, 433–441. https://doi.org/10.1016/j.tplants.2020.12.018

Groom, Q., Dillen, M., Addink, W., Ariño, A. H. H., Bölling, C., Bonnet, P., Cecchi, L., Ellwood, E. R., Figueira, R., Gagnier, P.-Y., Grace, O. M., Güntsch, A., Hardy, H., Huybrechts, P., Hyam, R., Joly, A. A. J., Kommineni, V. K., Larridon, I., Livermore, L., … Gaikwad, J. (2023). Envisaging a global infrastructure to exploit the potential of digitised collections. *Biodiversity Data Journal*, *11*, e109439. https://doi.org/10.3897/BDJ.11.e109439

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 19451–19452. https://doi.org/10.1073/pnas.1116812108

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE*, *6*, e19254. https://doi.org/10.1371/journal.pone.0019254

Hollingsworth, P. M., Li, D.-Z., van der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *371*, 20150338. https://doi.org/10.1098/rstb.2015.0338

Hughes, D. P., & Salathe, M. (2016). An open access repository of images on plant health to enable the development of mobile disease diagnostics. [WWW document] https://doi.org/10.48550/arXiv.1511.08060 [accessed 10 September 2024]

IUCN. (2025). The IUCN red list of threatened species. Version 2024–2. Retrieved 15 January 2025 from, https://www.iucnredlist.org

Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., Wagner, N. D., Emerson, B. C., Albach, D. C., Scheu, S., Bradler, S., de Vries, J., Irisarri, I., Li, H., Soltis, P., Mäder, P., & Wäldchen, J. (2024). Species delimitation 4.0: Integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution*, *39*, 771–784. https://doi.org/10.1016/j.tree.2023.11.002

Larridon, I. (2022). A linear classification of Cyperaceae. *Kew Bulletin*, *77*, 309–315. https://doi.org/10.1007/s12225-022-10010-x

Larridon, I., Rabarivola, L., Xanthos, M., & Muasya, A. M. (2019). Revision of the Afro-Madagascan genus *Costularia* (Schoeneae, Cyperaceae): Infrageneric relationships and species delimitation. *PeerJ*, *7*, e6528. https://doi.org/10.7717/peerj.6528

Larridon, I., Verboom, G. A., & Muasya, A. M. (2018). Revised delimitation of the genus *Tetraria*, nom. cons. prop. (Cyperaceae, tribe Schoeneae, Tricostularia clade). *South African Journal of Botany*, *118*, 18–22. https://doi.org/10.1016/j.sajb.2018.06.007

Larridon, I., Zuntini, A. R., Léveillé-Bourret, É., Barrett, R. L., Starr, J. R., Muasya, A. M., Villaverde, T., Bauters, K., Brewer, G. E., Bruhl, J. J., Costa, S. M., Elliott, T. L., Epitawalage, N., Escudero, M., Fairlie, I., Goetghebeur, P., Hipp, A. L., Jiménez-Mejías, P., Sabino Kikuchi, I. A. B., … Baker, W. J. (2021). A new classification of Cyperaceae (Poales) supported by phylogenomic data. *Journal of Systematics and Evolution*, *59*, 852–895. https://doi.org/10.1111/jse.12757

Le Bras, G., Pignal, M., Jeanson, M. L., Muller, S., Aupic, C., Carré, B., Flament, G., Gaudeul, M., Gonçalves, C., & Invernón, V. R. (2017). The French Muséum National d'Histoire Naturelle vascular plant herbarium collection dataset. *Scientific Data*, *4*, 170016. https://doi.org/10.1038/sdata.2017.16

Lee, S. H., Chan, C. S., & Remagnino, P. (2018). Multi-organ plant classification based on convolutional and recurrent neural networks. *IEEE Transactions on Image Processing*, *27*(9), 4287–4301. https://doi.org/10.1109/TIP.2018.2836321

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. [WWW document]. accessed 17 September 2024. https://doi.org/10.48550/arXiv.2103.14030

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. Retrieved 20 September 2024, from [WWW document] URL https://doi.org/10.48550/arXiv.1711.05101

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, *422*, 4694–4703. https://proceedings.neurips.cc/paper_files/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf

Park, J., de Lutio, R., Rappazzo, B., Ambrose, B., Michelangeli, F., Watson, K., Belongie, S., & Little, D. P. (2024). NAflora-1m: Continental-scale high-resolution fine-grained plant classification dataset. *Journal of Data-Centric Machine Learning Research*, *9*, 1–21. https://data.mlr.press/assets/pdf/v01-9.pdf

Parr, C., Wilson, N., Leary, P., Schulz, K., Lans, K., Walley, L., Hammock, J., Goddard, A., Rice, J., Studer, M., Holmes, J., & Corrigan, R. Jr. (2014). The encyclopedia of life v2: Providing global access to knowledge about life on Earth. *Biodiversity Data Journal*, *2*, e1079. https://doi.org/10.3897/BDJ.2.e1079

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, *721*, 8026–8037. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Pearline, A., & Kumar, S. (2019). DDLA: Dual deep learning architecture for classification of plant species. *IET Image Processing*, *13*, 2176–2182. https://doi.org/10.1049/iet-ipr.2019.0346

Perez, M. F., Bonatelli, I. A. S., Romeiro-Brito, M., Franco, F. F., Taylor, N. P., Zappi, D. C., & Moraes, E. M. (2022). Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. *Molecular Ecology Resources*, *22*, 1016–1028. https://doi.org/10.1111/1755-0998.13534

Pinheiro, F. d. C., Forzza, R. C., Leitman, P. M., & Prado, J. (2024). The REFLORA program: Implementation, repatriation, and creation of the REFLORA virtual herbarium as a tool for biodiversity studies. *Biota Neotropica*, *24*, e20241701. https://doi.org/10.1590/1676-0611-bn-2024-1701

Rasaminirina, F., Wieczorkowski, J. D., Rakotoarimanana, V., Rafaralahy, V. L., Rakotonirina, N., Ralimanana, H., & Larridon, I. (in press). Importance of habitat diversity in the Central Highlands for Cyperaceae conservation in Madagascar. Biotropica.

Richardson, J. E., Fay, M. F., Cronk, Q. C. B., Bowman, D., & Chase, M. W. (2000). A phylogenetic analysis of Rhamnaceae using *rbcL* and *trnL-F* plastid DNA sequences. *American Journal of Botany*, *87*, 1309–1324. https://doi.org/10.2307/2656724

Richardson, J. E., Fay, M. F., Cronk, Q. C. B., & Chase, M. W. (2000). A revision of the tribal classification of Rhamnaceae. *Kew Bulletin*, *55*, 311–340. https://doi.org/10.2307/4115645

Rzanny, M., Mäder, P., Deggelmann, A., Chen, M., & Wäldchen, J. (2019). Flowers, leaves or both? How to obtain suitable images for automated plant identification. *Plant Methods*, *15*, 77. https://doi.org/10.1186/s13007-019-0462-4

Rzanny, M., Wittich, H. C., Mäder, P., Deggelmann, A., Boho, D., & Wäldchen, J. (2022). Image-based automated recognition of 31 Poaceae species: The most relevant perspectives. *Frontiers in Plant Science*, *12*, 804140. https://doi.org/10.3389/fpls.2021.804140

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015. 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Seeland, M., & Mäder, P. (2021). Multi-view classification with convolutional neural networks. *PLoS ONE*, *16*, e0245230. https://doi.org/10.1371/journal.pone.0245230

Shirai, M., Takano, A., Kurosawa, T., Inoue, M., Tagane, S., Tanimoto, T., Koganeyama, T., Sato, H., Terasawa, T., Horie, T., Mandai, I., & Akihiro, T. (2022). Development of a system for the automated identification of herbarium specimens with high accuracy. *Scientific Reports*, *12*, 8066. https://doi.org/10.1038/s41598-022-11450-y

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Accessed 12 September 2024, from [WWW document] URL https://doi.org/10.48550/arXiv.1409.1556

Simpson, D. A., & Inglis, C. A. (2001). Cyperaceae of economic, ethnobotanical and horticultural importance: A checklist. *Kew Bulletin*, *56*, 257–360. https://doi.org/10.2307/4110962

Smith, V., Hardy, H., & Wainwright, T. (2022). DiSSCo UK: A new partnership to unlock the potential of 137 million UK-based specimens. *Biodiversity Information Science and Standards*, *6*, e91391. https://doi.org/10.3897/biss.6.91391

Söderkvist, O. J. O. (2001). Computer vision classification of leaves from Swedish trees. Master's Thesis, Linkoping University. https://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. Proceedings of the AAAI conference on artificial intelligence. Retrieved 5 October 2024 from, [WWW document]. https://doi.org/10.48550/arXiv.1602.07261

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 2016, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Tan, K. C., Liu, Y., Ambrose, B., Tulig, M., & Belongie, S. (2019). The herbarium challenge 2019 dataset. [WWW document] URL https://doi.org/10.48550/arXiv.1906.05372 [accessed 1 September 2024]

Thiers, B. M. (2024). Strengthening partnerships to safeguard the future of herbaria. *Diversity*, *16*, 36. https://doi.org/10.3390/d16010036

Wäldchen, J., Rzanny, M., Seeland, M., & Mäder, P. (2018). Automated plant species identification—Trends and future directions. *PLoS Computational Biology*, *14*, e1005993. https://doi.org/10.1371/journal.pcbi.1005993

White, E., Soltis, P. S., Soltis, D. E., & Guralnick, R. (2023). Quantifying error in occurrence data: Comparing the data quality of iNaturalist and digitized herbarium specimen data in flowering plant families of the southeastern United States. *PLOS ONE*, *18*(12), e0295298. https://doi.org/10.1371/journal.pone.0295298

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., & Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, *7*, e29715. https://doi.org/10.1371/journal.pone.0029715

Wu, S. G., Bao, F. S., Xu, E. Y., Wang, Y-X., Chang, Y-F., & Xiang, Q-L. (2007). A leaf recognition algorithm for plant classification using probabilistic neural network. 2007 IEEE International Symposium on Signal Processing and Information Technology, Giza, Egypt, 2007. 11–16. https://doi.org/10.1109/ISSPIT.2007.4458016

Xanthos, M., Mayo, S. J., & Larridon, I. (2023). Reassessment of morphological species delimitations in the *Cyperus margaritaceus-niveus* complex using morphometrics. *Plant Ecology and Evolution*, *156*, 112–127. https://doi.org/10.5091/plecevo.97453

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Arno, J., Morel, J., Rasaminirina, F., de Fátima Maciel-Silva, J., Cahen, D., Little, D. P., Silvestro, D., Antonelli, A., Grace, O., Zhang, L., & Larridon, I. (2025). A pipeline to compile expert-verified datasets of digitised herbarium specimens for automated plant identification to accelerate taxonomy. *Plants, People, Planet*, 1–15. https://doi.org/10.1002/ppp3.70149