Floralens: a Deep Learning Model for the Portuguese Native Flora

António Filgueiras[1], Eduardo R. B. Marques[1,2], Luís M. B. Lopes[1,2], Miguel Marques[1], Hugo Silva[1]

[1]Department of Computer Science
Faculty of Sciences, University of Porto
[2]CRACS/INESC-TEC

Abstract

Machine-learning techniques, namely deep convolutional neural networks, are pivotal for image-based identification of biological species in many Citizen Science platforms. However, the construction of critically sized and sampled datasets to train the networks and the choice of the network architectures itself remains little documented and, therefore, does not lend itself to be easily replicated. In this paper, we develop a streamlined methodology for building datasets for biological taxa from publicly available research-grade datasets and for deriving models from these datasets using off-the-shelf deep convolutional neural networks such as those provided by Google's AutoML Vision cloud service. Our case study is the Portuguese native flora, anchored in a high-quality dataset, provided by the Sociedade Portuguesa de Botânica, scaled up by adding sampled data from iNaturalist, Pl@ntNet, and Observation.org. We find that with a careful dataset design, off-the-shelf machine-learning cloud services produce accurate models with relatively little effort that rival those provided by state-of-the-art citizen science platforms. The best model we derived, dubbed Floralens, has been integrated into the public website of Project Biolens, where we gather models for other taxa as well. The dataset used to train the model and its namesake is publicly available on Zenodo.

Keywords: automatic identification, citizen science, deep learning, computer vision

1 Introduction

The improvements in processing speed, storage capacity, and imaging sensors for mobile devices paved the way for Citizen Science [1] applications

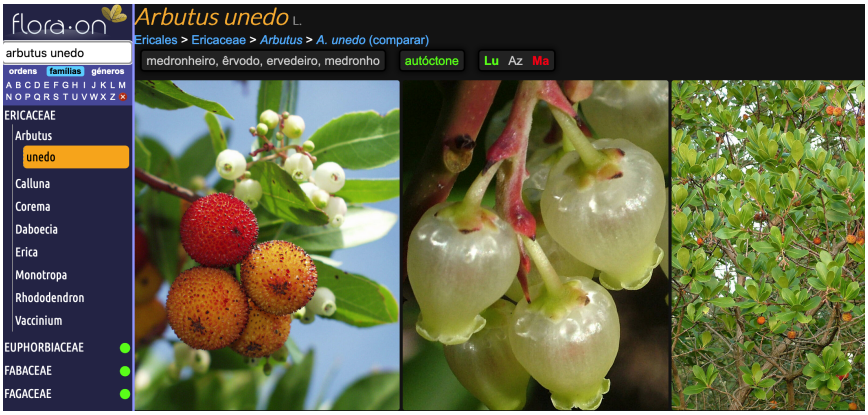


Figure 1: Detail of the FloraOn web application.

and Web services that allow amateur enthusiasts to participate in science projects. One very successful case study is that of nature observation, namely, the photographic recording of animals, plants, and fungi in their natural habitats. Besides storing these observations, some platforms use deep-learning models to provide automatic taxonomic identification from user-provided images [2, 3, 4, 5]. The data gathered by such projects is highly valuable for scientists, from hardcore taxonomists to ecologists studying the impact of human activity on biodiversity [6, 7].

This paper describes the step-by-step derivation of high-accuracy machine-learning models for automatic taxonomic identification of the Portuguese native flora. The work is anchored on the FloraOn dataset provided by the Sociedade Portuguesa de Botânica and available online via a web application[1] (Figure 1) and as a contributed dataset in the Global Biodiversity Information Facility (GBIF)[2]. While this dataset contains relatively few images per species, those provided are of very high quality, and all the identifications are provided by experienced taxonomists. We use the list of Portuguese native species in this dataset[3] as our reference to build a definite dataset that adequately covers all FloraOn species and allows accurate models to be derived using off-the-shelf convolutional neural networks (CNN) such as those provided by Google's AutoML Vision (GAMLV) cloud service.

While there are a few platforms that already provide automatic identification of flora species [2, 3, 8, 4], we wanted to develop a streamlined

[1] `https://flora-on.pt/`

[2] `https://www.gbif.org/`

[3] A few invasive or naturalized exotic species are also listed. They were not removed.

methodology to build a dataset for the Portuguese native flora and to derive an accurate model using off-the-shelf machine-learning tools. The goal would be to replicate the method for other biological taxa in the context of our ongoing Project Biolens [9]. The project aims to provide a set of CNN-based models for the taxonomic identification of biological species native to Portugal. Currently, we have four models: Floralens (described in this paper, covering the kingdom Plantae); Lepilens and Mothlens (for butterflies and moths, together covering the order Lepidoptera), and; Dragonlens (for dragonflies and damselflies, covering the order Odonata).

The methodology used to derive all these models shares two core traits with that described in this paper for Floralens: (a) the use of research-grade public repositories for dataset construction, and; (b) the use of GAML to derive the actual models. This methodology is succinctly described in a short scientific outreach article (in Portuguese) [10], and, more thoroughly, in MSc theses [11, 12] (both covering different stages of the work on Floralens), and a BSc project report [13] (covering Lepilens). Compared to the other models, Floralens was a greater challenge due to the much larger dimension of the domain, despite being limited to Portuguese native species.

We find that with a careful design of a custom dataset from publicly available research-grade datasets, current off-the-shelf machine-learning cloud-based services, such as GAMLV, produce impressive results with relatively little effort, even rivaling the results obtained with the aforementioned platforms. Thus, the main contributions of this work are as follows:

- a methodology to produce a dataset for a given biological taxa based on published research-grade datasets, e.g., from GBIF;
- a high-accuracy GAMLV-based model for the Portuguese native flora publicly available via web and mobile applications;
- a quantitative evaluation of the derived model and a comparison of its accuracy relative to state-of-the-art platforms such as Pl@ntNet;
- the complete dataset available on Zenodo.

The remainder of this paper is structured as follows. Section 2 describes the current state-of-the-art regarding automatic taxonomic identification based on deep learning. Section 3 describes the construction of the datasets used in this study. Section 4 describes the generation of the models from the datasets using GAMLV. Section 5 describes the results obtained with the models. Section 6 describes the software artifacts and datasets produced in

the scope of this work. Finally, Section 7 summarizes the main findings of this study and puts forward some future research goals.

2 Related Work

Convolutional Neural Networks (CNN) are deep neural networks composed of one or more layers of trainable convolutional nodes whose aggregate outputs are eventually fed to a final, fully connected, layer for tasks such as classification. In the context of image processing, a convolution is an operation that applies a matrix known as a *kernel* to a given input matrix. The kernel slides over the input, multiplying the overlapping matrix positions at each value and then summing the values. Depending on the form of the kernel, the resulting matrix can encode features such as edges, textures, and shapes, extracted from the original image.

The advent of CNN allowed the development of the first tools for the automatic identification of plant species from input images [14, 15, 16]. The success of these first efforts and their further refinement quickly reached a point in which automatic species identification rivaled identifications made by specialists [17], thus attesting to the transformative role of AI in this field [18]. Nowadays, models based on CNN are central tools in major citizen science platforms such as iNaturalist [2] and Observation.org [3] and Pl@ntNet [4]. The massive amount of image data, labeled by experts and/or crowd-sourcing efforts, enables the appearance of ML models that these platforms make available to users through web browser interfaces and/or mobile applications. Recently, Vision Transformers and their hybrids with CNN have emerged as new powerful tools to derive high-precision models for image classification tasks [19, 20].

When compared to the aforementioned citizen science platforms whose domain of application is the global flora, Floralens is more specialized covering only the Portuguese native species. It is also not supported by a citizen science platform nor, for the time being, supports directly exporting data to them, although it allows users to seamlessly export their observations into easily manageable data formats such as CSV files (and ZIP for images). In this respect, Floralens is closer in philosophy and implementation to the Flora Incognita Project [8].

Citizen science platforms generate important by-products in the form of their curated datasets, often made available to the public through biodiversity data portals, notably GBIF [21]. These datasets enable the development of other ML models as is the case of Floralens that, in addition to FloraOn,

uses data sampled from GBIF datasets provided by iNaturalist, Observation.org, and Pl@ntNet (cf. Section 3).

To improve the identification precision, some platforms such as Pl@ntNet developed regional models by dividing the globe into several biogeographic domains [22] following published regional floras such as the WCVP/Kew [23], and introduced metadata providing information on the anatomic part of the plant depicted in an image, e.g., flower, leaf, stem. This extra metadata critically improves the precision of the models [24, 25].

However, in our high-mobility world, many plants have escaped and significantly widened their native geographies, with or without mindful help from humans. The situation is such that many authors argue that specialized, regional models are not that useful. Powered by the high precision achieved in recent years by ML models, they focus instead on the generation of global models based on extreme datasets, consisting of millions of images representing tens of thousands of individual species (for comparison, it is estimated that the world has ∼300K plant species), as exemplified by the iNat Challenge [26] and PlantCLEF/LifeCLEF [27, 28].

3 Dataset Construction

We now describe the creation of the Floralens dataset, subsequently used for training the deep learning model. Our universe of species was taken from the FloraOn catalog as of November 2021, the reference point in time that marked the beginning of our work. It covers 2,712 species of native Portuguese flora. The dataset construction involved the identification of potential images of interest, forming a preliminary raw dataset, and then, through sampling, a characterization of the actual dataset that would be suitable for training a deep neural network, one in which each species is represented by a minimum of 50 and a maximum of 200 images. These bounds were defined based on our prior experience with building datasets for other biological taxa (cf. Section 6).

The FloraOn repository is composed of geo-referenced records of Flora species with associated images. The image data is relatively broad in scope, as it covers 78% of the entire catalog (2,127 out of 2,712 species), but has a limited volume: on average there are just 11 images per species, and, unsurprisingly, our 50-image lower bound threshold is not met for a single species. Hence, to adequately populate the Floralens dataset, we retrieved the FloraOn images but also consider image data from three publicly-available datasets stored and made available at GBIF by three citizen science plat-

forms: iNaturalist [29], Observation.org [30], and Pl@ntNet [31].

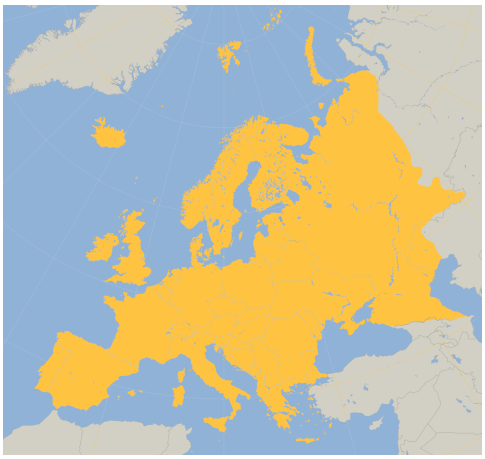


Figure 2: The geographical region of interest for GBIF portal queries.

The GBIF datasets provide validated observation data and associated images, originally submitted by users of the respective platforms. However, the validation process differs among data sources, as discussed further down in this section. For each dataset and for each species in our universe, we used GBIF portal queries to obtain observation records in the Darwin Core Archive format [32]. Each such record corresponds to an observation of a specimen, typically made in the wild by citizen scientists or experts, accompanied by its taxonomic identification, its geographical location, and one or more images. The GBIF portal queries were parameterized to cover the European continent (Figure 2) as, in a preliminary analysis, we found that occurrence data from just Portugal or even the entire Iberian Peninsula would yield limited data in terms of volume and variety. This was possible because, despite several endemisms, most species in the Portuguese native flora are widely distributed in the continent.

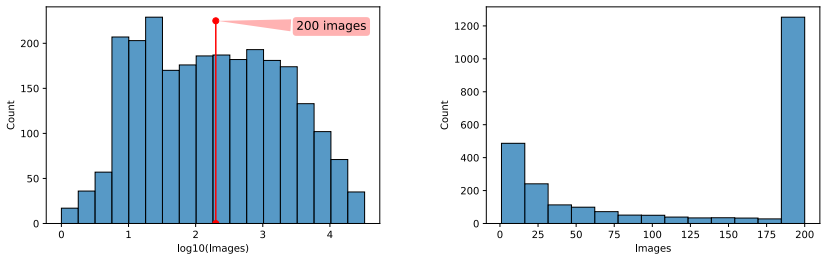
The raw data, from all image sources, is listed in Table 1 (left), along with the characterization of the Floralens dataset (right) that results from sampling the raw data. The corresponding histograms relating to image counts and the number of species are illustrated in Figure 3. In the raw data, more than 4 million images were available for consideration, covering 2,539 species (93% of the FloraOn catalog). Only 0.5% of these images are from FloraOn, and approximately two-thirds are taken from iNaturalist. Moreover, only 1,678 species reached our lower bound threshold of 50 images (61% of the FloraOn catalog). Given that most of the images are taken from

Table 1: Raw data and derived dataset after sampling (#I: image count; #S: species count; $\geq$ 50 I: species with more than 50 images).

	Raw data				Dataset		
Source	#I	%I	#S	$\geq$ 50 I	#I	%I	#S
FloraOn	22,869	0.5	2,127	0	15,191	5.1	1,397
iNaturalist	2,753,167	66.2	2,066	1,431	90,127	30.6	1,358
Observation.org	823,389	19.8	1,816	1,114	85,746	29.2	1,093
Pl@ntNet	515,950	12.4	1,495	735	102,537	34.9	1,373
Total	4,154,895		2,539	1,678	293,601		1,678

Citizen Science platforms, this scarcity can be due to subjective issues like the visual attractiveness of the plant, e.g., having a showy flower, or it can be a real effect, reflecting its rare status in the wild. The raw data distribution (in Figure 3a, shown in logarithmic scale) is long-tailed, in line with varied levels of abundance of species in nature.

The Floralens dataset was derived by sampling the raw data as follows. First, we filtered out species with less than 50 images. Then, for each of the remaining species, we sampled up to 200 images from the datasets, prioritizing data sources in the following order: (1) FloraOn; (2) Pl@ntNet; (3) Observation.org, and; (4) iNaturalist. That is, for the 50-200 image target per species, we use up as many images as possible from FloraOn first, then from Pl@ntNet, and so on.



(a) Raw data. (b) Floralens dataset.

Figure 3: Dataset histograms (x-axis: #images; y-axis: #species).

The intent of this source-based prioritization is to define a dataset where images are less prone to identification errors, taking into account the curation processes associated with each data source. The FloraOn data is

curated by botanic experts and images are of very high quality, typically clear images of specimens, often featuring subtle details that help secure the identification of a species. Pl@ntNet data goes through a curation process that involves machine learning, contributors' reputation scores, and geo-based species verification [31]. Observation.org data can result from automatic validation through image recognition coupled with a check for other approved observations in the geographical vicinity, or through an expert volunteer when automated validation fails [30, 33]. Finally, iNaturalist identifications result from a crowd-sourcing effort whereby "research-grade" identifications can be attained with the effort of a few, possibly just two and non-expert, citizen scientist labels [29, 34].

The end result of this process was the Floralens dataset, formed by approximately 300,000 images, and covering 1,678 species. As illustrated in Figure 3b, there are 200 images or very close to it for most of the species. The image count is 200 for 67% (1,128) of the species, 150 or higher for 79% (1,323), and 100 or higher for 86% (1,449). The data source prioritization scheme lead to a more significant fraction of FloraOn images in comparison to the raw data (the fraction grows from 0.5 to 5.1%) and, also, to a relatively even distribution of images from iNaturalist, Observation.org, and Pl@ntnet (the corresponding fractions are 30.6, 29.2, and 34.9%).

4 Model Derivation

The process of deriving an image classification model using GAMLV is illustrated in Figure 4. Overall, it comprises three stages: (1) preparing the data set for training; (2) training the model, and (3) deploying the model onto a cloud server or (using a suitable format) onto edge devices. GAMLV essentially requires the user to focus on the dataset preparation (1), given that training (2) and deployment (3) merely require simple high-level options by the user and are otherwise automated [35, 36]. The interaction with GAMLV can be conducted via a browser with a simple user interface, as we illustrate partially in this section (cf. Figure 5), or programmatically using Google Cloud APIs (e.g., in Python).

Data set preparation. This first step requires the user to load the dataset images onto what is called a storage bucket, provided by the Google Cloud Storage service (GCS), along with a simple CSV file characterizing the dataset. The CSV file lists the GCS image URIs and associates each URI to a ground truth label (the name of the species in the image) and to either

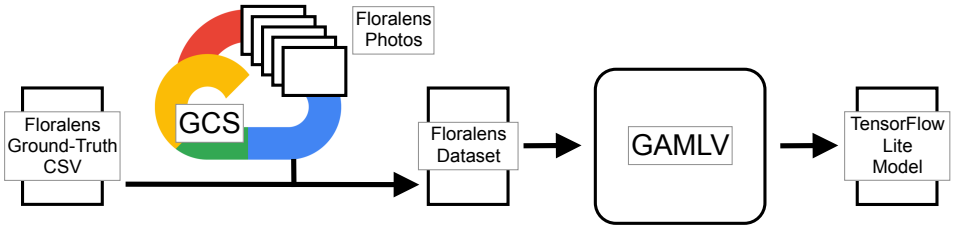


Figure 4: Model derivation using GAMLV.

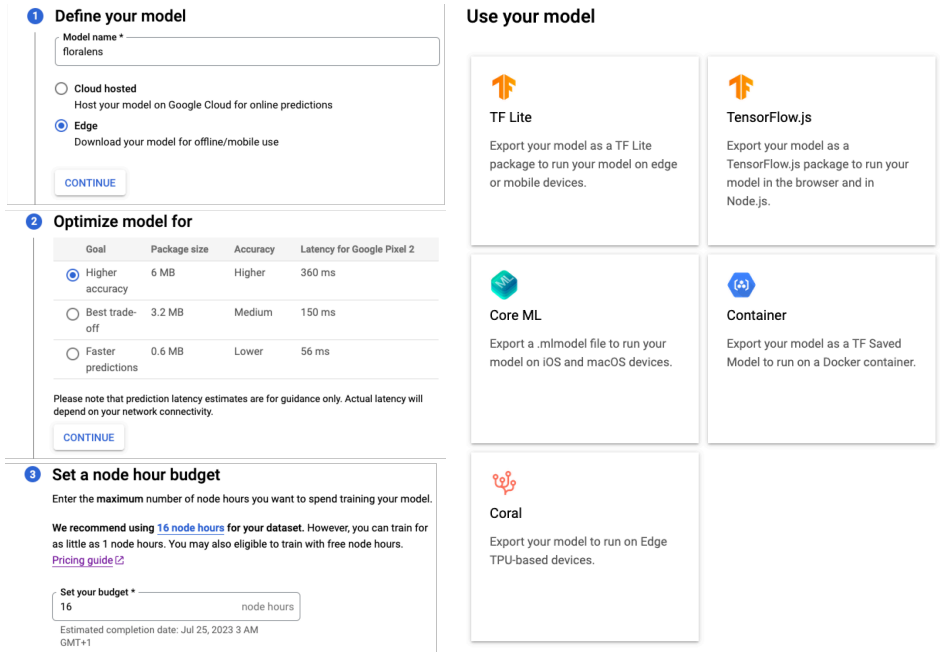
the train, validation, or test subset.

We fed GAMLV with train, validation, and test splits over the Floralens dataset, corresponding to fractions of 80%, 10%, and 10%. As usual, the train split is used to adjust CNN parameters during the training process in an iterative feedback loop, the validation split is used to measure the progress and convergence of that training process, and the test split is used merely for the evaluation of the model after training. The splits, with the image counts and data source provenance detailed in Table 2, resulted from a random selection of images for each species. Since the selection process is random and given the volume of images at stake, the overall fraction of images of each data source in each of the splits closely matches that of the overall dataset.

Table 2: Train, validation, and test splits over the Floralens dataset.

Data source	Train	Valid.	Test	Total
FloraOn	12,175	1,466	1,550	15,191 (5%)
iNaturalist	72,174	8,924	9,029	90,127 (31%)
Observation.org	68,614	8,603	8,529	85,746 (29%)
Pl@ntNet	81,918	10,367	10,252	102,537 (35%)
All	234,881 (80%)	29,360 (10%)	29,360 (10%)	293,601 (100%)

In [12], we consider other strategies for defining these splits. In particular, we explored approaches that gave preference to specific data sources for the validation/test splits. We found that a random split, besides preserving a roughly similar fraction of images per data source in each split, results in models with better performance (contrast the results in Section 5 with those in [12]). In any case, prioritizing particular data sources (e.g. FloraOn or Pl@ntnet) for validation/test splits over others had little impact on model performance.



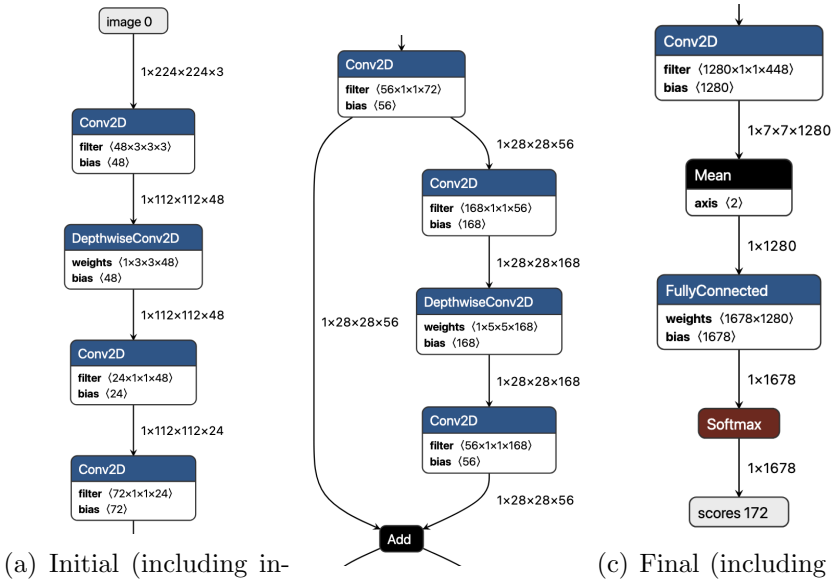
(a) Training parameters. (b) Deployment options.

Figure 5: GAMLV interface for model training and deployment.

Training. Once the dataset is imported onto AutoML, training may proceed, requiring only the user to make high-level choices for the type of model to be generated and the maximum training time, as illustrated in Figure 5a. In our case, we select the "edge" model option, given that we wish to host it as part of web or mobile applications (cf. Section 6) rather than deploying it in a Google Cloud server. We also toggle the option for a model which favors accuracy over latency among the three available choices. The maximum training time is specified in terms of a "node hours" budget, where nodes are virtual machines used during training.

GAMLV required 4 node hours to complete the training of the CNN with the Floralens dataset. It operates as a "black box" though, meaning that it is not possible to discern what goes on during training. For instance, no exact details or configuration options are provided for the training infrastructure (e.g., in terms of virtual machines, GPUs, or TPUs) and it is not possible to track details regarding the training process (e.g., how the model converges over time).

Deployment. Once training is completed, a model can be deployed in several formats, as illustrated in Figure 5b. The formats include the standard SavedModel format used by TensorFlow, but also others like TF Lite [37], an optimized TensorFlow format for use in mobile and embedded devices, or TFJS [38], for use in web browsers or Javascript programs. We make use of the TFLite and TFJS variants in the software artifacts described in Section 6. For these, GAMLV allows the user the export the associated files to a GCS bucket.



(a) Initial (including input layer).

(b) Intermediate.

(c) Final (including output layer).

Figure 6: Layers of the CNN model (fragment).

Derived CNN. The model obtained by GAMLV is a 65-layer deep CNN, with the structure partially illustrated in Figure 6 for the TF Lite version, in terms of the input/initial layers (a), intermediate layers (b), and final/output layers (c). The TFLite version differs from the standard TensorFlow model only in terms of post-training optimizations like quantization that enable the model to be interpreted faster with little degradation in accuracy [37]. As shown, the input layer takes a $224 \times 224 \times 3$ tensor, corresponding to a 224×224 (typically resized) image with 3 RGB channels, with 8-bit values per color channel. The intermediate layers make heavy use of stacked 2D convolutions and depth-wise convolutions with a repeating pattern. The

final layers include the derivation of a 1280-feature map that is fully connected to a soft-max activation function that produces the final classification vector with the label probabilities, 1678 of them in line with the number of species covered.

The CNN architectures at stake are picked from the MnasNet family [39], developed with mobile and embedded devices in mind. The high-level choice between models offered by GAMLV (back in Figure 5a) corresponds to three different MnasNet instantiations that do not differ in structure, just in the density of connections between layers.

5 Model Evaluation

In this section, we present an evaluation of the Floralens model using the test split described in the previous section, hereafter designated by FLTS (Floralens test split) using standard metrics. We then complement these baseline results with those obtained using two other test sets: a subset of the images in PlantCLEF'22-23 [40, 28] and a set of plant images automatically collected from Wikipedia. Furthermore, we compare the Floralens results for all test sets with Pl@ntnet models accessible through the Pl@ntNet API [41].

Evaluation Metrics. Precision is the ratio of true positives (TP) relative to the total number of positives (TP + FP). A positive (identification) occurs when the classification score (a probability) returned by the model equals or exceeds the established confidence level. Recall is the ratio of true positives relative to the total number of true examples (TP + FN).

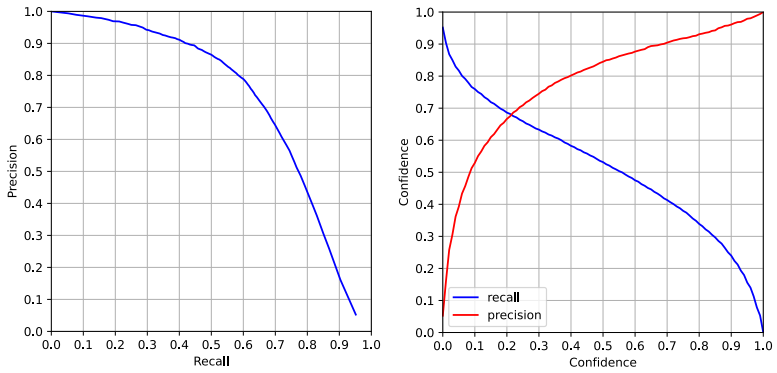
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Top-1 is the fraction of test images that the model correctly classified by the label with rank 1 (the highest-scoring label). Top-5 is similar to Top-1 but accounts for test images with a rank lower or equal to 5 (the 5 highest-scoring labels). We also use a variant of the Mean Reciprocal Rank (MRR) for test images of rank less or equal to 5. These are defined as follows:

$$Q(r_l) = \{t \in T \mid rank(t) \leq r_l\}$$

$$\text{Top-1} = \frac{|Q(1)|}{|T|} \qquad \text{Top-5} = \frac{|Q(5)|}{|T|} \qquad \text{MRR} = \frac{1}{|T|} \sum_{t \in Q(5)} \frac{1}{rank(t)}$$

where T is the set of all test images ($|T|$ = 29,360 as given in Table 2), $rank(t)$ is the rank of the ground truth label returned by the model for the test image t, and $Q(r_l)$ is the subset of T that contains test images with rank less or equal to a limit r_l.



(a) Precision-Recall curve. (b) Values per confidence level.

Figure 7: FLTS results: precision and recall results.

Baseline results. Figure 7 shows the results for precision and recall for the Floralens model applied to the test set given in Table 2, more precisely the macro-average of precision and recall values for all species to factor. The area-under-curve (AUC) for the precision-recall correlation (in 7a), also known as the average precision, is 0.72 (the maximum value would be 1.0). Putting the confidence levels in perspective (in 7b) we can visualize that precision and recall are both approximately equal to 0.7 for a confidence level of 0.2. For a confidence level of 0.5 precision equals 0.85 and recall equals 0.53. Overall, the results indicate a reasonable predictive power for the Floralens model.

Table 3: FLTS results: Top-1, Top-5 and MRR.

Data source	Top-1	Top-5	MRR
FloraOn	0.70	0.88	0.77
iNaturalist	0.70	0.87	0.77
Observation.org	0.64	0.83	0.72
Pl@ntNet	0.66	0.87	0.74
Overall	0.67	0.86	0.75

Table 3 lists the Top-1, Top-5 and MRR results for the FLTS, per data source used in the construction of the data set (Floralens, iNaturalist, Observation.org, and Pl@ntNet) and also in overall terms (last line in the table). The results indicate relatively homogeneous predictive power across all data sources, as the maximum difference in values for different data sources does not exceed 0.06: a Top-1 value of 0.64 for Observation.org vs. corresponding values of 0.7 for FloraOn and iNaturalist. The overall measures again indicate reasonably good predictive power: 0.67 for Top-1 (hence roughly two thirds of images in the FLTS are correctly classified with rank 1), 0.85 for Top-5, and 0.75 for MRR.

PlantCLEF and Wikipedia test sets. We consider two additional test sets: a random sample of 10,000 labeled images from the PlantCLEF'22-23 [40, 28] competition, and a sample of close to 1,500 images obtained from Wikipedia.

The PlantCLEF data we use is only a small sample of the entire "trusted" training set of PlantCLEF [42] that comprises approximately 2.9 million images covering 80,000 plant species. The repository is trusted in the sense that the image labels were obtained from academic sources or collaborative platforms like Pl@antNet or iNaturalist. Our subset was built by first filtering out species that are not covered by the Floralens model, obtaining data for 1593 (out of 1,678) species, and then randomly sampling 10,000 images.

As for the Wikipedia test set, the images were identified through the Wikimedia REST API search functionality [43]. For each species in the Floralens domain, we used the species name as the keyword for a REST API search. Among other items of information, the search result typically yields a reference to an image stored at Wikipedia which we then considered for addition to the test set. After obtaining the images, we filtered images that contained illustrations or herbarium specimens, as well as duplicate images (associated with more than one species; duplicates typically arise because the search may yield an image of a different species in the same genus if the target species' name does not have a Wikipedia page). Through this process, we obtained a dataset of 1,351 images for an equal number of species (one image per species). Compared to the PlantCLEF test set, the identification of the species in these images is less reliable as it results from an uncontrolled crowd-sourced effort with no specific directives for image validation.

Table 4 shows the results of the Floralens model for these test sets considered in terms of the Top-1, Top-5, and MRR metrics. We also recall the

overall FTLS results (from Table 3) for easy comparison. We can observe that the PlantCLEF and Wikipedia test set results are marginally lower than those obtained for the FLTS, (by 0.02/0.03 in all metrics). Overall, the results are evidence that the Floralens model applies well to other datasets beyond the base test suite.

Table 4: Top-1, Top-5, and MRR of the Floralens model for all test sets.

Dataset	#I	#S	Top-1	Top-5	MRR
FLTS	29,360	1,678	0.67	0.86	0.75
PlantCLEF	10,000	1,593	0.65	0.84	0.73
Wikipedia	1,351	1,351	0.65	0.84	0.72

Genus results. The results for the genus model (Table 5) show a clear overall improvement relative to the species model. The greatest enhancement is observed for the Wikipedia test set and, especially, for the Top-1 result ($\Delta = +0.14$). The latter is probably because while this test set is less exact than the others, when the image on the Web page of a species is wrongly labeled Wikipedia does manage to provide an image of a plant of the same genus. The improvements observed for FLTS and PlantCLEF for all metrics are the same.

Table 5: Top-1, Top-5, and MRR of Floralens for genus prediction (Δ: variation relative to species results).

Dataset	Top-1	Δ	Top-5	Δ	MRR	Δ
FLTS	0.76	+0.09	0.91	+0.05	0.82	+0.07
PlantCLEF	0.74	+0.09	0.89	+0.05	0.80	+0.07
Wikipedia	0.79	+0.14	0.91	+0.07	0.83	+0.08

Comparative Pl@ntNet API results. We now provide results comparing the Floralens model with models accessible via the Pl@ntNet API [41]. The Pl@ntNet API is a RESTful web service that provides access to the same visual identification models used by state-of-the-art Pl@ntNet apps [4]. The API lets us obtain a set of ranked species for a given image for two models for worldwide flora: a so-called "legacy" model from 2022, henceforth identified as PN^{22}, generated using CNN, and; a recent model announced in July

Table 6: Pl@ntNet API: comparative MRR values (Δ: variation relative to the Floralens model).

Dataset	PN^{22}	Δ	PN^{23}	Δ	PN^{23F}	Δ
FLTS	0.68	−0.07	0.80	+0.05	0.80	+0.05
PlantCLEF	0.72	−0.01	0.79	+0.06	0.79	+0.06
Wikipedia	0.73	+0.01	0.78	+0.06	0.79	+0.07

Table 7: Floralens vs Pl@ntNet API: MRR per data source in the FLTS (Δ: variation relative to the Floralens model).

Source	PN^{22}	Δ	PN^{23}	Δ	PN^{23F}	Δ
FloraOn	0.58	−0.17	0.79	+0.04	0.79	+0.04
iNaturalist	0.67	−0.10	0.81	+0.04	0.81	+0.04
Observation.org	0.59	−0.13	0.76	+0.04	0.76	+0.04
Pl@ntNet	0.77	+0.03	0.84	+0.10	0.84	+0.10
FLTS \ Pl@ntNet	0.63	−0.12	0.78	+0.03	0.78	+0.03
FLTS	0.68	−0.07	0.80	+0.05	0.80	+0.05

2023 [44], generated using Vision Transformers, henceforth PN^{23}. Through the API, it is also possible to filter results from the PN^{23} model so that only species occurring in a specific biogeographic region are included. One of these regions is Southwestern Europe which includes Portugal, allowing the most head-to-head comparison between Floralens and Pl@ntNet that can be devised. These results are identified by PN^{23F}.

Table 6 shows the variation (Δ) of the MRR values obtained for PN^{22}, PN^{23} and PN^{23F} for all the test sets relative to the corresponding values obtained for Floralens. PN^{22} performs worse than Floralens for the FLTS, a variation of −0.07. The MRR values of PN^{22} are otherwise similar for PlantCLEF (−0.01) and Wikipedia (+0.01). The discrepancy observed for FLTS merits further analysis and is discussed below. Focusing now on PN^{23} and PN^{23F}, the MRR values across all test sets range from 0.78 to just 0.80, and perform better than Floralens by a factor of 0.05 to 0.07. The Southwestern Europe species filter associated with PN^{23F} has little impact on the results.

In Table 7 we show the results in more detail for the FLTS by discriminating the data sources. The goal is to understand why Floralens shows better results than PN^{22} and also the impact of Pl@ntNet images in the

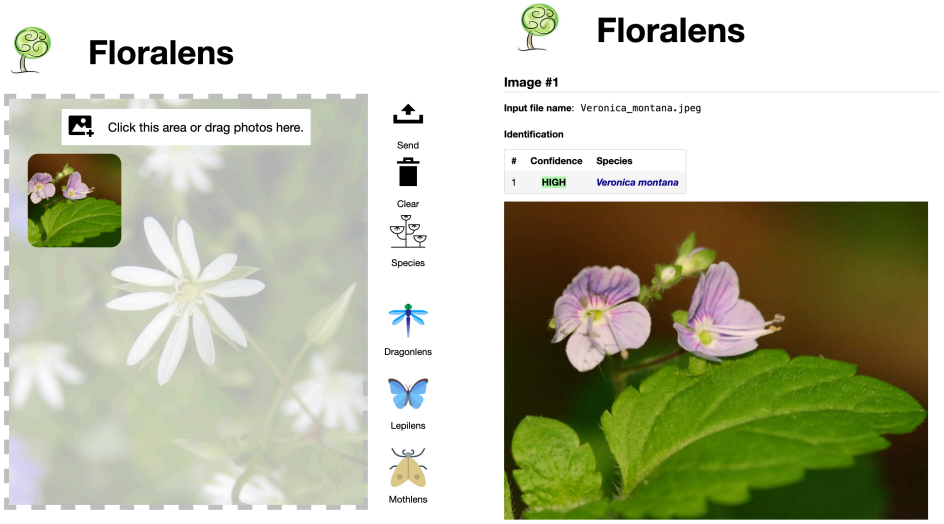
MRR values. Recall that Pl@ntNet data was used to define our model. That is, of course, also the case for Pl@ntNet models. In particular, part of the Pl@ntNet data we use for testing may have been used to train the Pl@ntNet models. That could explain the fact that the MRR values are noticeably higher for the Pl@ntNet test subset (row Pl@ntNet in Table 7) when compared to the remaining test suite overall (row FLTS \ Pl@ntNet in Table 7). This effect is clearer in the case of PN^{22} (0.77 vs. 0.63) but, also, in the case of PN^{23} and PN^{23F} (0.84 vs. 0.78). PN^{22} has $\Delta = -0.07$ for FLTS and the value goes down to $\Delta = -0.12$ when we exclude Pl@ntNet images from FLTS. Subject to the same restriction, PN^{23}/PN^{23F} have MRR values of 0.80 ($\Delta = +0.05$) versus 0.78 ($\Delta = +0.03$), respectively, both corresponding to modest improvements relative to Floralens.

Overall, the Floralens results are on par and in some cases better than PN^{22}, and marginally worse than PN^{23} and PN^{23F}.

6 Software Artifacts

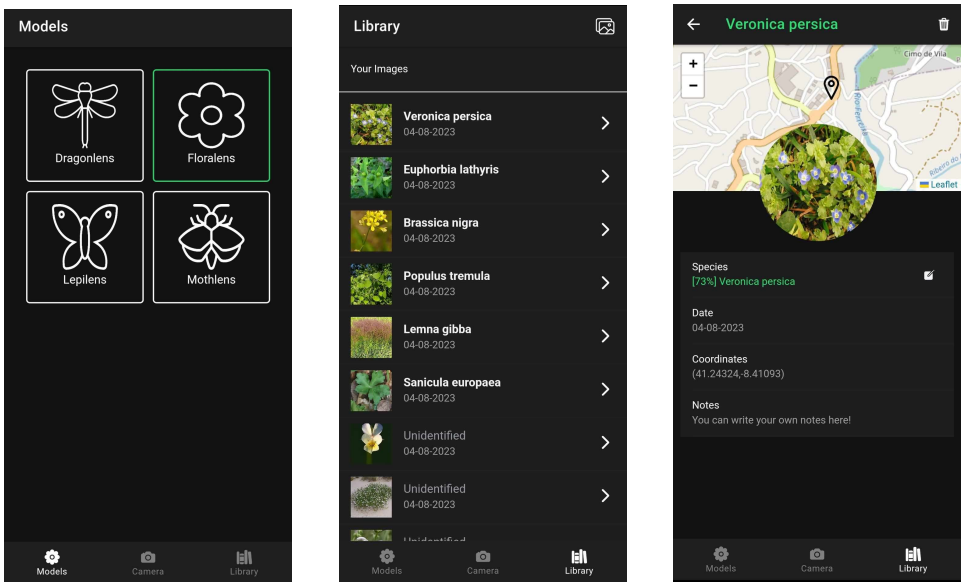
Biolens web site. The Floralens model has been integrated into the Biolens project website [9]. The functionality is quite simple: users submit photos of interest and obtain corresponding suggestions of biological identifications, as illustrated in the screenshots of Figure 8. To enable this deployment, the Biolens website is hosted by a small virtual machine that requires just 2 CPU cores and 8 GB of RAM. The configuration is quite lightweight, given that we make use of the TFLite variant of the Floralens model (and similarly for other models hosted on the site).

Biolens Android app. We also recently developed a prototype version of a mobile application that can run on Android and iOS devices. The Android version is available for download at the Biolens website. A few screenshots of the application are shown in Figure 9. The functionality is similar to that of the Biolens website, but customized for a mobile application context: users can take photos of specimens on the fly and obtain instant identification suggestions without an Internet connection. This information, together with the date, the current geographical location, and optional user annotations, is recorded in association with each photo. Another important aspect is that the app can be used without Internet access. All Biolens models are bundled within the app and, thus, are evaluated *in loco* on the mobile device.



(a) Image uploading by the user. (b) Suggested automatic identification.

Figure 8: Biolens – web application screenshots.



(a) Model selection. (b) Image list. (c) Details for an image.

Figure 9: Biolens – mobile application screenshots.

Floralens dataset and results. Finally, the full Floralens dataset is publicly available on Zenodo [45]. The dataset contains the mapping between the image labels (ground truth), the image URLs from which they were retrieved, URLs for a site we maintain where all images are also stored, and GBIF identifiers when applicable (all images except those obtained from FloraOn). Ground truth and URLs are also available for the PlantCLEF and Wikipedia datasets used in the evaluation of Section 5, as well as for all datasets the top-5 results and corresponding confidence levels for the Floralens model and the three Pl@ntNet model variations.

7 Conclusions

In this paper, we present the methodology used in the construction of the Floralens dataset for the Portuguese native flora and in the derivation of a deep-learning model for the automatic identification of the species therein. The universe of species was taken from the FloraOn dataset, provided by the Sociedade Portuguesa de Botânica and compiled exclusively by specialists. The dataset was constructed based on high-quality data from several research-grade datasets available via GBIF. Besides FloraOn these include: iNaturalist, Pl@ntNet, and Observation.org. We made the dataset available to the community on Zenodo [45]. The Floralens model was derived from this dataset using GAMLV, a platform that provides users with tools to derive models from datasets using off-the-shelf convolutional deep neural networks.

Our initial tests suggested that the Floralens model had good predictive power, with an AUC metric value of 0.72 and, for a reference confidence level of 0.5, values for precision and recall of 0.85 and 0.53, respectively. Further experiments indicated a relatively homogeneous predictive power across all data sources used in the dataset, with a maximum variation of 0.06, and they confirmed its good predictive power with values for Top-1 and Top-5 of 0.67 and 0.86, respectively. Compared with the state-of-the-art platform Pl@ntnet, Floralens performed on par with the "legacy 2022" model and only marginally worse when compared with the most recent one. We integrated the model into the BioLens Project website and developed a mobile application to allow using it offline, in the field.

As for future work, we aim to improve the species coverage and the accuracy of the model. One way to do that is to include data from other datasets such as those of Encyclopedia of Life [46] and FloraIncognita [8]. We also want to address some limitations of the dataset that arise from the inconsis-

tent use of taxonomic names and synonyms. Our list of Portuguese native species taken from FloraOn is as complete and up-to-date as possible. However, recent taxonomic revisions have changed the accepted binomial names for some species and these adjustments are not immediately reflected in the public datasets. As an example, the species featured in the FloraOn listing as *Atractylis gummifera* is now known as *Chamaeleon gummifer*. Although not very common, it is widespread in the Mediterranean region [47]. Nevertheless, it is not included in the Floralens dataset as we did not find enough (≥ 50) images with our GBIF queries for *Atractylis gummifera*. However, a recent query for images of *Chamaeleon gummifer* yields more than enough images to include the species in the Floralens dataset in a future update.

More work is also required on the Biolens mobile app to improve its usability and optimize resource usage. Integration with existing Citizen Science platforms is a possibility, allowing the user to automatically upload Biolens records.

Finally, we plan to continue preliminary work on developing classification models from image similarity analysis based on the TensorFlow Similarity package [12]. Such models provide an alternative way to clinch an identification when GAMLV-based models yield low-confidence results. Hybrid classification models that combine both approaches are an interesting possibility.

Acknowledgements. The authors would like to thank Hugo Gresse, Pierre Bonnet, and Mathias Chouet from Project Pl@ntNet for kindly providing us with extended access to the Pl@ntNet API for our analysis. This work was partially funded by projects SafeCities and Augmanity (POCI-01-0247-FEDER-041435 and -046103, through COMPETE 2020 and Portugal 2020), and by project UIDB/50014/2020 (Fundação para a Ciência e Tecnologia). It would not have been possible without the support from the Google Cloud Research Credits program.

References

[1] R. Bonney, C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience*, 59(11):977–984, December 2009.

[2] S. Altrudi. Connecting to nature through tech? The case of the iNaturalist app. *Convergence*, 27(1):124–141, 2021.

[3] M. Schermer and L. Hogeweg. Supporting citizen scientists with automatic species identification using deep learning image recognition models. *Biodiversity Information Science and Standards*, 2018.

[4] A. Affouard, H. Goëau, P. Bonnet, J. C. Lombardo, and A. Joly. Pl@ntNet app in the era of deep learning. In *International Conference on Learning Representations*, Toulon, France, 2017.

[5] J. Wäldchen and P. Mäder. Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11):2216–2225, 2018.

[6] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(1):792, 2022.

[7] S. Christin, E. Hervet, and N. Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 2019.

[8] P. Mäder, David Boho, Michael Rzanny, Marco Seeland, Hans Christian Wittich, Alice Deggelmann, and Jana Wäldchen. The Flora Incognita app–interactive plant species identification. *Methods in Ecology and Evolution*, 12(7):1335–1342, 2021.

[9] Biolens. `https://rubisco.dcc.fc.up.pt/biolens`. Accessed September 2022.

[10] L. M. B. Lopes, E. R. B. Marques, T. Mamede, A. Filgueiras, M. Marques, and M. Coutinho. Identificação taxonómica em biologia usando inteligência artificial. *Revista de Ciência Elementar - Casa das Ciências*, December 2022.

[11] M. Marques. A Portuguese Flora Identification Tool Using Deep Learning. Master's thesis, Masters thesis, Faculty of Sciences, University of Porto, 2021.

[12] A. Filgueiras. Floralens: a deep learning model for portuguese flora. Master's thesis, Masters thesis, Faculty of Sciences, University of Porto, 2022.

[13] T. Mamede. On using Deep Learning for Automatic Taxonomic Identification of Butterflies. Master's thesis, BSC project report, Faculty of Sciences, University of Porto, 2020.

[14] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino. Deep-plant: Plant identification with convolutional neural networks. In *IEEE International Conference on Image Processing*, pages 452–456, 2015.

[15] I. Heredia. Large-scale plant classification with deep neural networks. In *Computing Frontiers Conference*, pages 259–262, 2017.

[16] Y. Sun, Y. Liu, G. Wang, and H. Zhang. Deep learning for plant identification in natural environment. *Computational Intelligence and Neuroscience*, 2017.

[17] P. Bonnet, Hervé Goëau, Siang Thye Hang, Mario Lasseck, Milan Šulc, Valéry Malécot, Philippe Jauzein, Jean-Claude Melet, Christian You, and Alexis Joly. Plant identification: Experts vs. machines in the era of deep learning: deep learning techniques challenge flora experts. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pages 131–149, 2018.

[18] AI Nature Services. `https://ainature.eu/`. Accessed September 2022.

[19] A. Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, Virtual Conference, 2021.

[20] Y. Bai, J. Mei, A. L. Yuille, and C. Xie. Are Transformers more robust than CNNs? In *Advances in Neural Information Processing Systems*, pages 26831–26843, 2021.

[21] T. Robertson, M. Döring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, and P. Desmet. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the Internet. *PLOS One*, 9(8), 2014.

[22] A. Justamante, A. Joly, J. C. Lombardo, F. Robert, M. Chouet, S. Liñán, K. Soacha, and J. Piera. AI-GeoSpecies: integrate artificial intelligence into your citizen science app. `https://doi.org/10.5281/zenodo.7657594`, February 2023.

[23] WCVP: World Checklist of Vascular Plants. `http://sftp.kew.org/pub/data-repositories/WCVP/`. Accessed September 2023.

[24] M. Rzanny, M. Seeland, J. Wäldchen, and P. Mäder. Acquiring and preprocessing leaf images for automated plant identification: understanding the tradeoff between effort and information gain. *Plant Methods*, 13(1):1–11, 2017.

[25] M. Rzanny, P. Mäder, A. Deggelmann, M. Chen, and J. Wäldchen. Flowers, leaves or both? how to obtain suitable images for automated plant identification. *Plant Methods*, 15(1):1–11, 2019.

[26] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The iNaturalist species classification and detection dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[27] H. Goëau, P. Bonnet, and A. Joly. Plant identification based on noisy web data: the amazing performance of deep learning (LifeCLEF 2017). In *Conference and Labs of the Evaluation Forum*, 2017.

[28] H. Goëau, P. Bonnet, and A. Joly. Overview of PlantCLEF 2022: Image-based plant identification at global scale. In *Conference and Labs of the Evaluation Forum*, volume 3180, pages 1916–1928, 2022.

[29] iNaturalist contributors, iNaturalist (2022). iNaturalist research-grade observations. iNaturalist.org. `https://doi.org/10.15468/ab3s5x`. Accessed via GBIF.org on July 2023.

[30] H. de Vries and M. Lemmens. Observation.org, nature data from around the world. `https://doi.org/10.15468/5nilie`. Accessed via GBIF.org on July 2023.

[31] A. Affouard, A. Joly, J. C. Lombardo, J. Champ, H. Goeau, and P. Bonnet. Pl@ntnet observations. Version 1.2. Pl@ntNet. `https://doi.org/10.15468/gtebaa`. Accessed via GBIF on July 2023.

[32] T. Robertson, M. Döring, R. Guralnick, D. Bloom, J. Wieczorek, K. Braak, J. Otegui, L. Russell, and P. Desmet. The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PLOS One*, 9(8), 2014.

[33] Validation. `https://observation.org/pages/validation/`. Accessed July 2023.

[34] iNaturalist. What is the data quality assessment and how do observations qualify to become "research grade"? https://www.inaturalist.org/pages/help#quality. Accessed July 2023.

[35] E. Bisong. *Google AutoML: Cloud Vision*, pages 581–598. Apress, 2019.

[36] AutoML Vision Documentation. https://cloud.google.com/vision/automl/docs/. Accessed July 2023.

[37] TensorFlow Lite, ML for Mobile and Edge Devices. https://www.tensorflow.org/lite/. Accessed July 2023.

[38] TensorFlow.js, Machine Learning for Javascript developers. https://www.tensorflow.org/js/. Accessed July 2023.

[39] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2815–2823, 2019.

[40] Plantclef2022, image-based plant identification at global scale. https://www.imageclef.org/PlantCLEF2022. Accessed July 2023.

[41] Pl@ntNet API for developers. https://my.plantnet.org. Accessed July 2023.

[42] PlantCLEF'22 trusted training set. https://lab.plantnet.org/LifeCLEF/PlantCLEF2022/train. Accessed July 2023.

[43] Wikimedia REST API. Accessed July 2023.

[44] Pl@ntNet. Pl@ntNet news – Covering all countries floras and new identification AI. https://plantnet.org/en/2023/07/05/covering-all-countries-floras-new-identification-ai/. Accessed September 2023.

[45] The Floralens Dataset for Portuguese Flora. https://doi.org/10.5281/zenodo.10639701. Accessed February 2024.

[46] Encyclopedia Of Life Datasets. https://opendata.eol.org/dataset. Accessed November 2023.

[47] Royal Botanical Gardens, Kew: Plants of the World Online: Chamaeleon gummifer. https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:192416-1. Accessed July 2023.