Multimedia Tools and Applications
https://doi.org/10.1007/s11042-025-20883-2

A database for automatic identification of herbarium specimens in Piperaceae family

Alexandre Yuji Kajihara[1] · George Azevedo de Queiroz[2] · Marcelo Galeazzi Caxambú[3] · Luiz Eduardo S. Oliveira[4] · Diego Bertolini[1] · André Luis Schwerz[1]

Received: 7 March 2024 / Revised: 29 January 2025 / Accepted: 24 April 2025

Abstract
Herbaria worldwide have been digitizing their collections to preserve specimens, improve research access, and support biodiversity conservation amid environmental changes. This recent digitization process has revealed that thousands of plants have yet to be appropriately identified or reviewed because of the complex and time-consuming classification and the relatively low number of qualified expert taxonomists. Computer Vision techniques could be promising alternatives for supporting plant identification; however, there is a lack of criteria for designing reliable and representative datasets needed to develop robust classification systems. This often occurs because existing datasets aggregate multiple taxonomic groups with substantial differences among their species, failing to represent the practical realities of plant identification tasks in herbaria. To address this challenge, this work introduces a new database of herbarium specimens exclusively from the Piperaceae Giseke family, accompanied by a series of experiments conducted on this dataset. The Piperaceae, also known as the pepper family, is a large botanical family with many species that are intrinsically complex to identify due to their similarities. We selected 10,503 specimens samples on the *species*Link repository of 236 Piperaceae species across three genera, collected in Brazil. A comprehensive set of experiments evaluated segmentation, feature extraction, and classification algorithms for the dataset performance as reference values. The best performance combined non-handcrafted features (VGG16 and ViT) and the Multilayer Perceptron classifier. The difficulty in identifying some Piperaceae species is due to their morphological characteristics, which requires that the task be submitted for final review by an expert. We hope the database and our experiments described in this work will benefit the research community.

Keywords Machine Learning · Digitized herbarium specimen · Automated identification · Piperaceae · Plant species identification · Deep learning

Extended author information available on the last page of the article

Published online: 15 May 2025

1 Introduction

Over the last five centuries, herbaria have spread across the world thanks to the hard work of scientists. According to the 2021 Index Herbarium, organized by the New York Botanical Garden (USA), there are 3,522 active herbaria in the world, which store around 400 million plants and fungi [1]. While specimens deposited in herbaria may not always receive an immediate species-level identification, all they have documented morphological characteristics and typically include metadata such as the collector's name, collection date, and place of collection [2]. Therefore, these millions of specimens, carefully preserved over the centuries, are a unique source of information to help understand how the world's vegetation has changed over time and predict how it will change in the future [3].

Due to the ongoing accelerated destruction of biodiversity, thousands of species are in danger of extinction before being discovered, described, and identified. Public policies for conservation are fundamentally dependent on efforts to catalog the diversity of plants on Earth [4]. Thus, a worldwide effort has begun to document the planet's biodiversity in the last century. Most of the world's herbaria are currently involved in the process of digitizing their collections. Around 81 million specimens have already been digitized [3] and shared online through various open data repositories, such as the Global Biodiversity Information Facility (GBIF), the world's largest biodiversity open-data network [5]; and Integrated Digitized Biocollections (iDigBio), which makes data and images available from more than 1,600 US biological collections [6].

In Brazil, the digitization of specimens from herbaria started in 2010, and the images and metadata have been available in the Reflora Virtual Herbarium [7] and *species*Link [8]. The cooperative and collaborative network *species*Link aims to integrate and openly make available biological collections from Brazil and abroad [9]. In July 2023, the *species*Link had more than 17 million records of primary biodiversity data for more than 12 million botanical specimens, of which at least 4 million included images [8]. Figure 1 shows a digitized herbarium specimen used in this work. A dried and pressed plant fixed on a large sheet with a label containing information about the specimen, the collector name, date, and collection place is also named exsiccatae [10].

Herbarium online collections, including high-quality images and metadata, are relevant scientific contributions. However, new uses of this information essentially depend on the reliable identification of specimens [11]. This is still a sore point because thousands of plants need to be identified or have incorrect identification and, therefore, must be re-identified [12]. The large number of specimens deposited in herbaria awaiting identification is due to several factors: manual identification and classification workflows, which are slow and subject to errors [13]; and a limited number of new taxonomists in recent times [14]. Moreover, many current experts devote their careers to interpreting incomplete or imprecise descriptions [15]. These difficulties consume a substantial portion of systematic research today, redirecting efforts away from advancing the field. As a result, plant identification skills are currently confined to a relatively small group of individuals [15].

Computer vision techniques offer potential support in biological research by narrowing the gap in specimen identification. However, the main state-of-the-art contributions are often built upon large datasets encompassing species from a wide range of taxonomic groups. These species have significant morphological differences, including a high interclass diversity within the dataset, which may be easily recognized by specialists and effec-

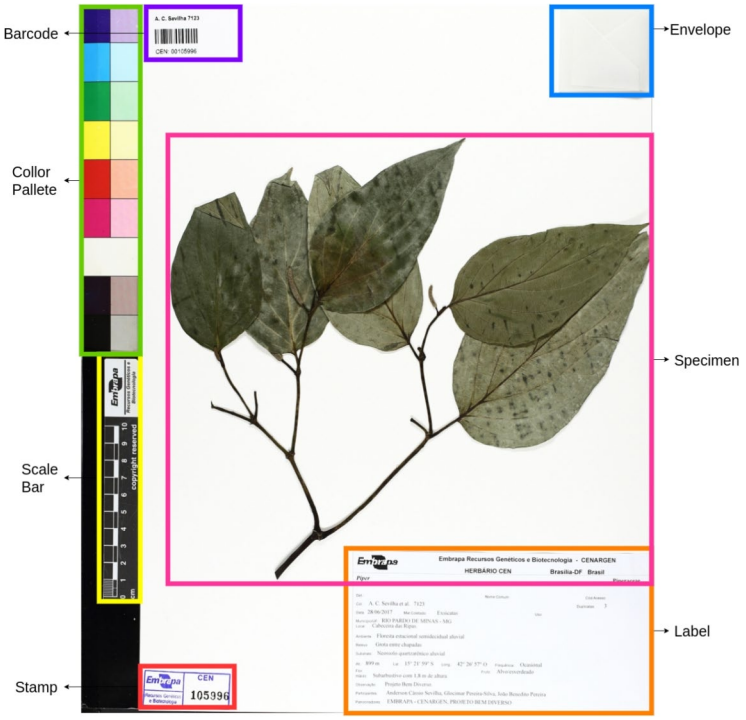


Fig. 1 Example of *Piper crassinervium* from Herbário da Embrapa Recursos Genéticos e Biotecnologia (CEN) [8], having the seven artifacts: scale bar, barcode, envelope, specimen, color pallet, stamp, and specimen label

tively captured by machine learning algorithms. Furthermore, these datasets often include species with many samples, diverging from the practical realities of plant identification tasks in herbaria.

The Piperaceae Giseke, commonly known as the pepper family, demands specialized expertise due to the vast number of existing species and their subtle distinguishing features [16]. Several studies highlight the inherent complexity of identifying Piperaceae species, especially within its largest genera: *Peperomia* Ruiz & Pavon (approximately 1,700 species) [17, 18] and *Piper* Linnaeus (around 2,600 species) [19]. The large volume of species [20] and high morphological similarity among many pose a significant challenge in Piperaceae specimen identification [21]. As far as we know, no efforts have yet addressed the task of species-level classification for herbarium specimens within the Piperaceae family.

The main contributions of our work are summarized as follows:

- A curated dataset of herbarium specimens of Piperaceae Giseke is a botanical family with inherently complex identification, comprising hundreds of species. He brings together plants collected in several Brazilian regions with 236 species and 10,503 images. The original database, segmented images, extracted features, and its different subsets are available for download at https://zenodo.org/records/14599766. As commonly found in herbarium collections, many species in this dataset have few samples, and few

species have many samples;
- A robust experimental protocol encompassing segmentation, feature extraction, and classification. Our protocol evaluates the dataset, focusing on comparing established classification models. We explore diverse subsets, handcrafted and non-handcrafted features, and different performance metrics. The performance of these experiments establishes a baseline for future research.

The rest of this paper is organized as follows. Section 2 shows the related work for classifying and segmenting herbarium specimens. The dataset built and refined for this study is presented in Section 3. The proposed approach for segmentation, feature extraction, and classification is detailed in Section 4. Section 5 describes the experimental results and discussion. Section 6 presents the conclusion and future work.

2 Related work

In this section, we discuss some important works described in the literature that have led to and contributed to our research. Although numerous studies use leaves collected in the field, we specifically concentrated on plants deposited in herbaria. Herbarium plants are dried and stored to preserve their characteristics over the years. The drying process alters the plant's cellular structure, causing a loss of color but preserving its essential traits required for identification. Plants are typically identified after drying since collection sites are unsuitable due to the absence of structure and unfavorable environmental conditions. Recently, we have seen works employing a cross-domain approach, using models trained from herbarium samples to identify plants collected in the field [22]. A comprehensive systematic review of herbaria-related tasks using Computer Vision and Machine Learning is presented in [23].

Given the diversity of proposed datasets covering different taxa, comparing performance between different studies is impossible. In this context, the main contribution of this work lies in providing a comprehensive database of a botanical family with inherently complex identification. Next, we present some of the most relevant herbarium datasets used in the literature. Furthermore, we discuss several Machine Learning approaches that have successfully segmented and classified herbarium specimens.

Datasets Previous studies on identification involving herbaria specimens have employed datasets with a reduced number of classes and samples [11, 24–29]. Clark et al. [24] employed four species of the genus *Tilia* and 516 samples of isolated leaves (129 of each species). Wijesinghe and Marikar [25] used 158 isolated leaves from 17 species of trees of the genus *Stemonoporus* (Dipterocarpaceae family). Grimm et al. [26] classified six fern species with 108 samples of isolated leaves. Unger et al. [27] classified 26 species of the most common trees in Germany with 260 examples. Kho et al. [28] applied 54 leaf image samples of three species of *Ficus*, one of the largest genera in the plant kingdom. Pryer et al. [11] classified 108 samples of three species of horsetails: *Equisetum hyemale*, *E. laevigatum*, and *E.* × *ferrissii*. Finally, Kajihara et al. [29] assembled a dataset with five genera of the Piperaceae family represented by 375 examples.

More recently, some datasets with large numbers of species (classes) and herbarium images (samples) have also been proposed and evaluated [12, 30–35]. Wilf et al. [30] assembled a dataset comprising 7,597 carefully curated images of cleared leaves (or leaflets for compound leaves), encompassing nearly an equivalent number of species, spanning 2,001 genera. Carranza-Rojas et al. [12] proposed two datasets: Herbarium255 (255 species with approximately 11,071 images) and Herbarium1K (1,204 species with 253,733 images). Schuettpelz et al. [31] assembled a dataset with two closely related families: 9,276 club-mosses (Lycopodiaceae) and 9,113 spikemosses (Selaginellaceae), in which, unlike the others, the target class is the family. Younis et al. [32] classified 1,000 species from 830,408 herbarium images. The Herbarium 2019 dataset proposed in [33] contains 46,469 digitally imaged herbarium sheets representing 683 species from the flowering plant. The Herbarium 2021 dataset is the largest ever proposed, containing 2.5 million images of vascular plant specimens representing approximately 64,500 taxa [34]. Shirai et al. [35] assembled a dataset with 500,554 specimen images of 2,171 plant taxa that grow in Japan.

A common characteristic in these larger datasets is gathering species from different botanical families. Consequently, the more significant number of species from many families and genera leads to high interclass variability. The only exception is Tan et al. [33], which selected samples of 683 species from the family Melastomataceae.

Machine Learning Studies on herbarium specimens in the last two decades have employed different classifiers. Grimm et al. [26], Kho et al. [28], Unger et al. [27], Wilf et al. [30], and Kajihara et al. [29] used Support Vector Machine (SVM) for classification of specimens. Other classifiers, such as Decision Tree (DT) and k-Nearest Neighbors (k-NN), were also used for the classification of herbarium specimens by Pryer et al. [11] and Kajihara et al. [29]. Neural Networks also are used in [24, 25, 28, 29]. Clark et al. [24] and Kajihara et al. [29] used Multilayer Perceptrons (MLP). Marikar [25] used a Probabilistic Neural Network (PNN).

Some works began employing Convolutional Neural Networks (CNNs) for specimen classification. Schuettpelz et al. [31] built a CNN for their study. Carranza-Rojas et al. [12] used a modified version of GoogLeNet. Carranza-Rojas et al. [36] extended previous work by including three CNNs specially built for the proposed work. Younis et al. [32] used a modified version of ResNet. Little et al. [2] employed modified ResNet, SeResNeXt, and SENet versions. Shirai et al. [35] used versions of Inception-ResNet, Inception, and VGG16. The VGG16 was also used by Pryer et al. [11] to classify herbarium specimens. The top-five teams used the GENet, ECA-NFNet-L0, and variations of the ResNet and ResNeXt architectures in the competition described in [37].

Segmentation Herbarium images typically feature various artifacts, such as stamps, labels, color palettes, envelopes, etc. In order to mitigate the impact of these artifacts on classification, it is essential to remove them. In smaller datasets [24–29], artifacts were manually removed, leaving only the plant or its leaves isolated. These artifacts were manually blurred in [33] and disregarded in [31, 35]. All images were uniformly cropped to remove the barcodes and notes on the specimen in [12, 32, 36]. It is worth noting that Shirai et al. [35] clarified that the presence of a label, color bar, scale, or stamp in the image did not significantly affect the identification accuracy in their dataset. On the other hand, some works have

been dedicated exclusively to segmenting herbarium specimens. White et al. [38] employed a modified version of U-Net to segment plant tissue in exsiccata. Triki et al. [39] proposed a network structure inspired in VGG16 for segmenting leaves and exsiccata artifacts. A combination of YOLO-V8 and U-Net++ for segmentation is addressed in [40]. Milleville et al. [41] proposed a pipeline utilizing YOLO-V8 and Mask R-CNN for detection and segmentation in exsiccata containing multiple specimens. Others works have dedicated efforts for component detection on digitized herbarium specimens. Younis et al. [42] detect plant organs with Faster R-CNN. The original YOLO-V3 and improved YOLO-V3 models were used by Triki et al. [43] for detection of components in specimens. Tompson et al. [44] used YOLO-V5 for detection of herbarium specimen sheet components.

Piperaceae-related works The only work for the identification task of herbarium specimens using the Piperaceae Giseke family was conducted by Kajihara et al. [29]. This study generated a pre-processed and balanced subset of data from the *species*Link repository. Its performance in identifying specimens from this family at the genus level was evaluated using 375 images. To evaluate which combination of descriptor and classifier would produce the best accuracy in the identification task, after pre-processing, Kajihara et al. [29] extracted features using Local Binary Pattern (LBP), Speed Up Robust Features (SURF), VGG16, ResNet50V2, and MobileNetV2. The data were used to train SVM, MLP, *k*-NN, and DT classifiers. The best accuracy was 80.53%, obtained with the combination of MobileNetV2 and SVM.

Unlike the works cited above, Pravin and Deepa [45] classified living plants in nature. Despite this, the work deserves attention because it dealt with species from the largest genus of the Piperaceae family called *Piper* L. Their dataset has 1,607 images of 15 *Piper* L. species, with approximately 100 samples for each species. Among their experiments, they achieved an F1-Score of 0.87 by combining the Random Forest (RF) classifier with features extracted by representation learning.

2.1 Critical review

The recent digitization process of specimen collections from herbaria allowed the introduction of Computer Vision and Machine Learning techniques to help herbaria curators classify specimens, mitigating dependence on specialists. On the other hand, the identification of herbarium specimens is still a challenging task. These challenges usually concern the inherent characteristics of the dataset, as discussed below.

Most datasets [11, 24–29] are balanced. However, herbarium collections are typically unbalanced due to several factors: some species are more readily found in nature than others [36]; in certain species, a particular plant organ, such as the fruit, is essential for precise species identification may be unavailable for collection during certain seasons of the year [28]; and older herbaria tend to have more samples of certain species due to their earlier start in collecting and storing specimens [28]. Imbalanced datasets [12, 30, 33–35] better represent herbaria but often exhibit a long-tailed data distribution, which may potentially affect overall classification performance.

The studies [12, 32, 35, 37] are among the main contributions to the identification of herbarium specimens. Their datasets have been incorporating an ever-greater number of spe-

cies to increase the number of samples. Consequently, they often include distantly related species with distinct morphological traits, which simplifies classification tasks for specialists in the real world. In contrast, our work focuses on a single taxonomic group, known among taxonomists for its complex identification due to the morphological similarity shared by many of its species.

As the automatic identification task is still new, there is no consensus on metrics to measure the performance of identification of herbarium specimens, making it challenging to compare new approaches with state-of-the-art works and build new benchmarks. As pointed out in [23], many studies use accuracy or mean reciprocal rank for species identification and average precision for segmentation tasks.

Digitized specimen images have substantial variations ranging from the position of the artifacts (such as scale bar, stamp, color pallet, and so on) to how the plant is fixed on the sheet. A challenging task is removing irrelevant parts of the image, keeping only the plant specimen in order not to bias the identification. Few works have dealt with the segmentation of digitized specimen images, but their training dataset is relatively small, with a limited set of species [23, 46].

3 Dataset

There are several herbarium plant specimen databases [11, 12, 24–28, 30–32, 34–36]. However, these databases usually include either samples of a limited number of species within a specific family or many samples of diverse families. Classification models trained on a variety of species but from different botanical families do not effectively support botanists and experts. This is because classifying specimens from different families is a relatively straightforward task. Taxonomists highlight that the real challenge in identification arises when dealing with species from the same family.

Labeling samples of dried plants (exsiccata) requires botanists' expertise in the respective plant families. For certain species, these experts, also known as taxonomists, must have physical samples (exsiccata) on hand to perform the classification. Consequently, transporting these physical samples, necessary for conducting evaluations, leads to associated costs and delays in the identification process. These practices underscore the importance of databases labeled by experts in the automated classification process.

In this paper, we selected from *species*Link [8] the digitized herbarium specimen images of the Piperaceae family, their metadata, such as the species name in Latin, identifiers, and the information contained in the labels of the specimens. In 2021, *species*Link had 52,606 records of Piperaceae. These specimens were collected in different Brazilian regions. Brazil has a vast territory with diverse climatic zones, varying altitudes and landforms, as well as distinct soil properties. This diversity contributes significantly to the wide variety of Piperaceae species found within its territory.

Despite the wide variety of records, many of them have missing images or just images of plants photographed in a natural environment instead of being dehydrated, missing data on the collection's origin, and incomplete or unreliable identification. To solve these inconsistencies, we perform extensive pre-processing to remove records without metadata about the identifier name, collection location, and complete botanical taxonomy (family, genus, and species). We also kept samples that contained herbarium images of dried plants.

The quality of species determinations is a weakness in the records maintained by *speciesLink*. To mitigate bias in classification due to mislabeled samples, we have chosen images of plants identified by 11 botanists recognized as specialists in the Piperaceae family. This task was supervised by a researcher dedicated to studying Piperaceae at the Rio de Janeiro Botanical Garden Research Institute (Brazil).

Figure 2 illustrates all steps made to build the dataset. We also only selected species that had at least five samples. As a result, our dataset contains 10,503 images of specimens of 236 species of Piperaceae collected in Brazil. More details about the dataset produced can be found at Zenodo[1].

Figure 3 illustrates the specimens' taxonomic distribution by family, genus, and species. Among 10,503 images of Piperaceae, 7,678 (73.10%) are of the genus *Piper* L., 2,806 (26.72%) of the genus *Peperomia* Ruiz & Pav., and 19 (0.18%) of the genus *Manekia* Trelease. There are 156 species of *Piper* L., 79 of *Peperomia* Ruiz & Pav., and only one species of *Manekia* Trel.

Figure 4 illustrates the distribution of the 10,503 collected samples (specimens) across the 236 classes of interest (species). The distribution reveals a set with long-tailed data, where few samples represent some species, while others have a large number of samples. Among these, there are 60 majority classes and 176 minority classes. A class is considered a majority when its number of samples exceeds the number of samples in a class in a balanced dataset [48].

The chosen samples comprise plants collected from various regions of Brazil. They are presently housed in 35 Brazilian herbaria and three herbaria located in the USA. Notably, The New York Botanical Garden (NY) herbarium holds the largest collection, consisting of

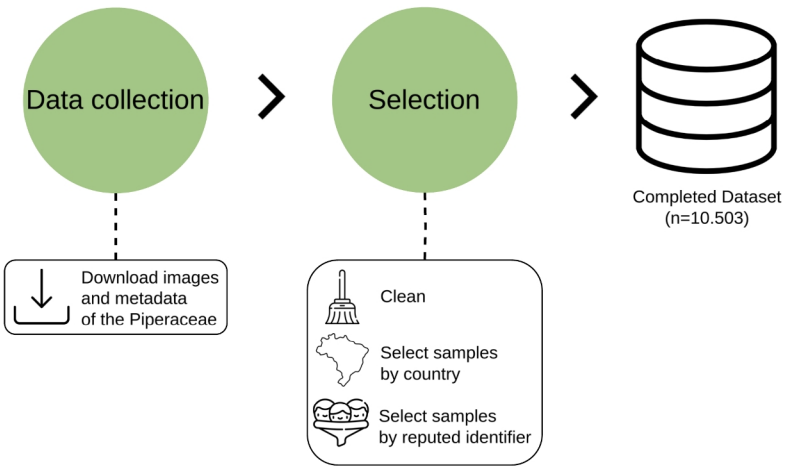


Fig. 2 Illustrative workflow to assemble the dataset of digitized images of specimens of the Piperaceae family from Brazilian herbaria

[1] https://doi.org/10.5281/zenodo.14599766

Fig. 3 A stacked pie chart made with Krona Tools [47] depicting the dataset taxonomic distribution by family, genus, and species

3,757 samples. Figure 5 depicts the distribution of the 10,503 samples across these diverse herbaria. The complete list of herbarium names is available on Zenodo[2].

In order to explore the dataset according to the collection areas, we propose a subset containing Piperaceae exsiccatae collected in the Paraná state. Figure 6(a) depicts the distribution of samples throughout the Brazilian states. Paraná has the highest average rate of images per species. We also divided the dataset into five Brazilian regions (North, Northeast, Midwest, South, and Southeast), as shown in Fig. 6(b). We can note that the Southeast region and Paraná state have the highest number of images for each species.

As previously mentioned, each class contains a minimum of 5 samples. However, there is a significant disparity in the number of images between minority and majority classes. To address this imbalance within the dataset, we divided the subsets into smaller fragments based on the minimum number of images per species. Consequently, we generated new subsets comprising species with a minimum of 5, 10, and 20 images.

[2] https://doi.org/10.5281/zenodo.14599766

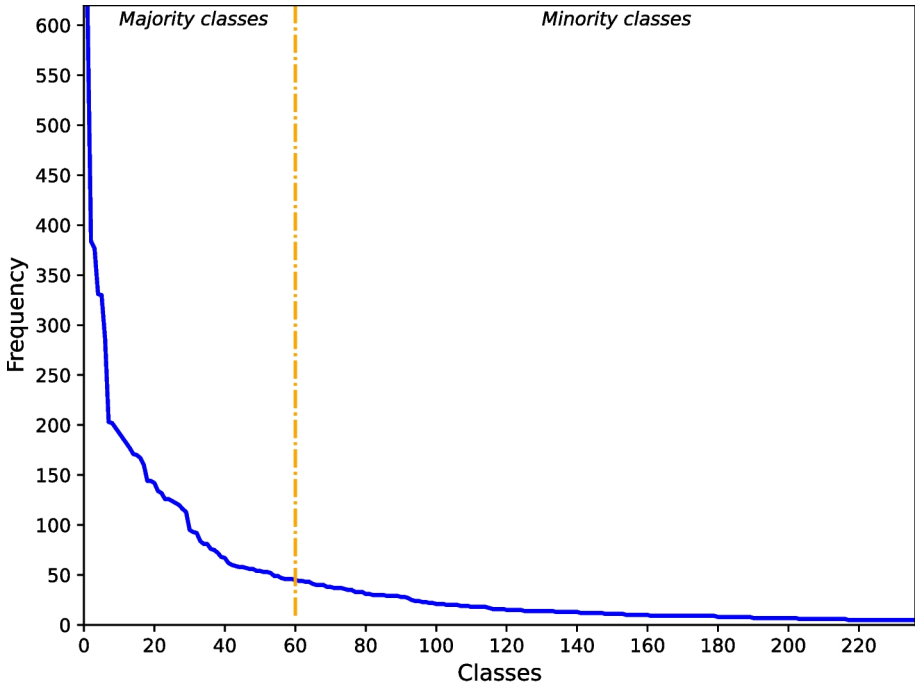


Fig. 4 Distribution of classes (species)

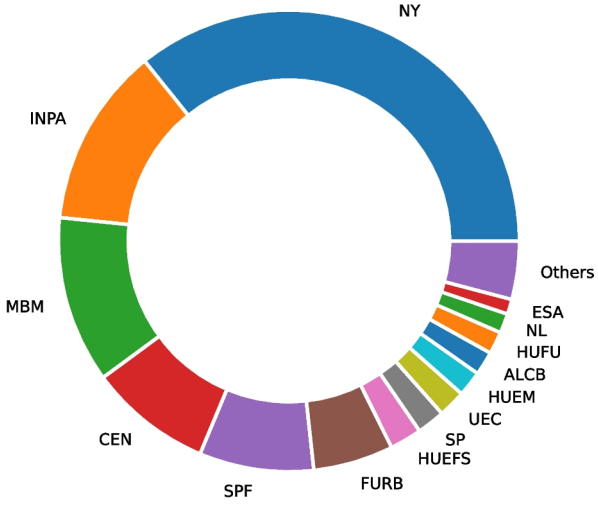
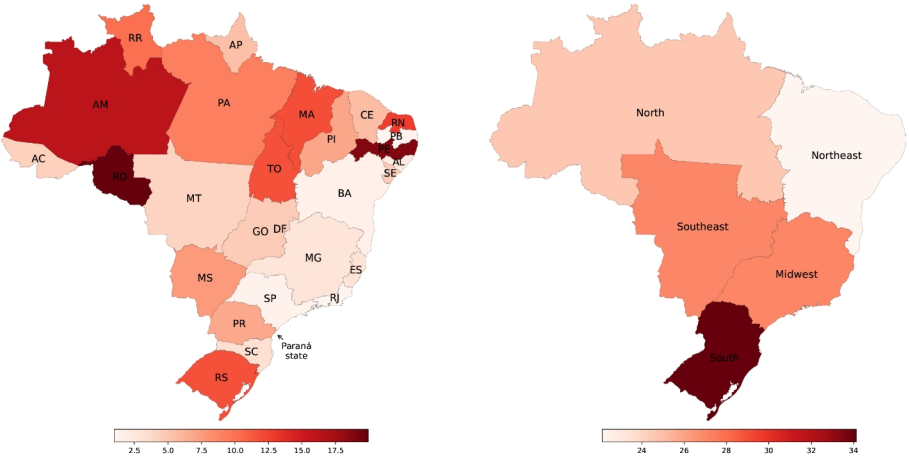


Fig. 5 Distribution of samples by herbaria. Each herbarium is represented by its acronym

Table 1 summarizes the number of images, the number of species, and the average rate of images per species for each new subset (5, 10, and 20) of the Paraná dataset. Similarly, Table 2 describes the same fragmentation for each region of Brazil. Moreover, we incorporate the imbalance degree for each subset, which illustrates the disparity between a hypothetical balanced distribution and the actual imbalanced subset. Imbalance degree is single



(a) States (b) Regions

Fig. 6 Geographical distribution of herbarium specimens of the Piperaceae family: (**a**) average of images per species in the states and (**b**) average of images per species in the regions

Table 1 Paraná subsets according to the minimum number of samples per species

# Minimum samples	Samples	Species/classes	Imbalance Degree
5	1,354	55	35.20
10	1,235	36	22.19
20	1,065	23	13.18

Table 2 Regional subsets according to the minimum number of samples per species

Regional subsets	# Minimum samples	Samples	Species/classes	Imbalance Degree
North	5	2,634	107	73.16
	10	2,351	68	41.17
	20	2,015	41	27.15
Northeast	5	1,062	48	30.18
	10	974	35	20.18
	20	776	21	12.16
Midwest	5	1,136	42	31.28
	10	1,044	29	20.29
	20	888	17	11.31
South	5	2,457	72	52.20
	10	2,309	49	32.20
	20	2,109	33	20.20
Southeast	5	2,756	102	66.17
	10	2,524	67	46.15
	20	2,211	42	25.16

Table 3 Complete dataset (Brazil) according to the minimum number of samples per species

# Minimum samples	Samples	Species/classes	Imbalance Degree
5	10,503	236	175.14
10	9,977	160	119.14
20	9,237	106	73.15

real value in the range $[0, K)$, where K is the number of classes. The closer the imbalance degree value is to zero, the closer the dataset is to a balanced distribution. This measure effectively distinguishes class distributions and exhibits a stronger correlation with the challenges posed by skewed class distributions in supervised algorithms [48]. A reduction in the degree of imbalance is observed when we restrict the minimum number of samples per class.

Table 3 describes the number of species, the number of samples, and the imbalance degree of the fragments produced for the complete dataset. It is worth noting that some images and species are only in the complete dataset but absent in the regional subsets. This absence occurs because the number of available images of a given species is insufficient for its inclusion in a regional subset. For example, there are five images of the species *Piper cunninghamii*, four of which originated in the North region and one in the Northeast region. These images are insufficient to include this species in the regional sets but are enough to compose the Brazilian dataset.

This section provided a detailed description of the curated dataset assembled from Piperaceae specimens collected in Brazil. Although there are species with only one or two collected samples, we adopted a minimum sample threshold of five to accommodate the five-fold cross-validation scheme. This strategy enabled the selection of 236 classes with 10,503 samples, which are available for download from the Zenodo repository[3].

4 Proposed method

The pipeline for identification of Piperaceae family specimens consists of three main steps: segmentation, feature extraction, and classification, as illustrated in Fig. 7. The source code is available on GitHub[4]. We describe below the details of each of them.

4.1 Segmentation

Usually, digitized images of herbarium exsiccatae contain several artifacts. These artifacts include stamps, labels, color palettes, envelopes, etc. The presence of these artifacts may introduce noise in the classification, and their positions vary from image to image. For this reason, the segmentation of dried plants is essential to avoid data unrelated to the plant during feature extraction.

In this paper, we use the U-Net architecture for semantic segmentation, which was initially proposed in [49] to handle biomedical images. The U-Net is a Fully Convolutional Network composed of the encoder, the bottleneck module, and the decoder. The U-Net has a U-shaped structure combined with context information; fast training speed; and good

[3] https://doi.org/10.5281/zenodo.14599766

[4] https://github.com/xaaaandao/piperaceae-identification-paper/tree/multimedia

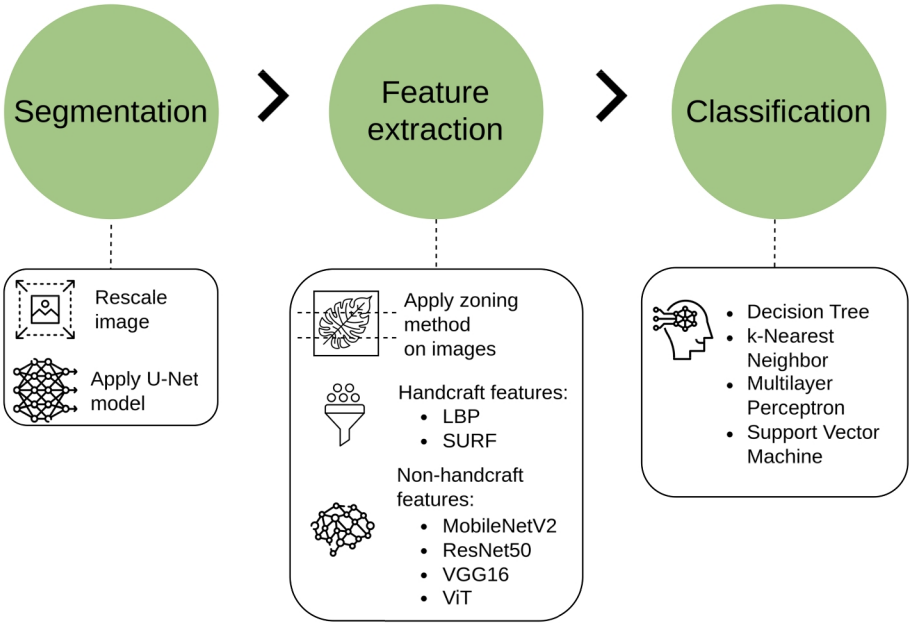


Fig. 7 Proposed method general scheme

Table 4 Main hyperparameters used in the U-Net

Hyperparameter	Value
Batch	4
Epoch	75
Learning rate	0.001

performance, even with a small amount of data annotated [50]. Such features meet medical image segmentation requirements but have also been used for herbarium image segmentation in [38, 39].

In this paper, we conducted training on the U-Net using a small dataset of Piperaceae specimens proposed in [29]. This dataset consists of 375 manually segmented images. We also evaluated the network's performance for image variations, including different color modes (RGB and grayscale) and dimensions (256×256, 400×400, and 512×512 pixels). These dimensions have also been evaluated to reduce the computational cost spent on training time and memory consumption in similar works [12, 31, 51].

Table 4 shows the main hyperparameters used in the U-Net. We implemented a Python script[5] using the scikit-learn[6] to split the folds, tensorflow[7] to create the model, and PIL[8] to save the segmented image. All digitized herbarium specimens were resized before perform-

[5] https://github.com/xaaaandao/piperaceae-identification-paper/tree/multimedia

[6] https://scikit-learn.org/stable/index.html

[7] https://www.tensorflow.org/api_docs/python/tf

[8] https://pillow.readthedocs.io/en/stable/

Table 5 List of hyperparameters for each extractor of features

Descriptor	Hyperparameter	Dimension
SURF	$SurfSize = 64$	257
LBP	$P = 8\,and\,R = 2 \times 11 - 8bit$	59
VGG16	As done by Simonyan and Zisserman [52]	512
ResNet50	As done by He et al. [53]	2048
MobileNet	As done by Howard et al. [54]	1280
ViT	As done by Dosovitskiy et al. [55]	1024

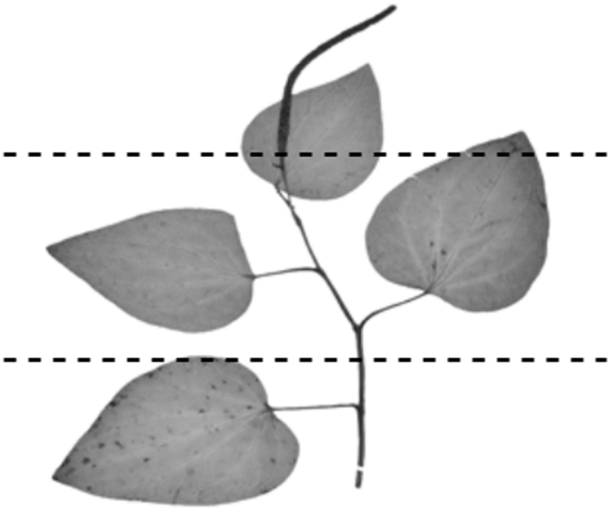


Fig. 8 Image split horizontally into three regions. Original exsiccata from [8]

ing the inference with the trained U-Net model. Then, having the dataset segmented by the U-Net, the feature extraction can be started as described below.

4.2 Feature extraction

The features of the Piperaceae family specimens were extracted using both handcrafted and learned feature descriptors. As handcrafted approaches, we employed the LBP and SURF descriptors; the other three CNNs (MobileNetV2, ResNet50, and VGG16) and ViT as non-handcrafted descriptors. Table 5 presents the parameters used and the sizes of the feature vectors generated by each extractor.

We included the image zoning method when we applied the feature extractors based on Representation Learning - MobileNetV2, ResNet50V2, VGG16, and ViT - on our dataset. We opted for the horizontal orientation and the division into three regions following previous results [29]. An example of this zoning is illustrated in Fig. 8.

Local Binary Patterns (LBP) [56] is a robust descriptor that effectively captures binary patterns in a texture. It is known for its ease of implementation and low computational complexity [57]. It compares the intensity of the central pixel with its eight neighboring pixels and assigns 0 to the neighbor when its value is less than that of the central pixel, and 1, otherwise [58]. These binary values are then multiplied by the weights given to the corresponding pixels. Afterward, the values of the eight pixels in the neighborhood are added, and the result is the number of that texture unit [56].

The Speeded Up Robust Features (SURF) [59] is a point-of-interest detector and descriptor for images with good run-time performance and speed, allowing real-time applications. Its main characteristic is the repeatability that permits finding the same points of interest in different visualization conditions [59].

To our knowledge, no studies have used LBP and SURF for feature extraction from dried herbarium plants. Despite this, the efficiency and robustness of LBP and SURF have been applied in [60, 61] with images of live leaves (photographed in a natural environment) to identify plants.

In addition to handcrafted features, this work extracted deep features through CNNs. The option for this type of neural network is because they have provided a notable advance in automated identification [23]. The three CNN models used in this work to extract deep features were MobileNet, ResNet50, and VGG16.

The MobileNet neural network [54] is designed for mobile applications and embedded systems such as object detection, face recognition, and large-scale geolocation. Unlike traditional CNN, where the convolution layer filters and combines the inputs into a new set of outputs, the MobileNet uses depth-wise separable convolutions in which one layer filters and another later combines the outputs. This reduces the network size and its computational complexity [54].

ResNet is a CNN developed by He et al. [53] to solve a problem observed in deep neural networks of accuracy saturation, followed by rapid degradation, which occurs with increasing network depth. To address this issue, the ResNet is built with residual blocks to send the gradients directly to the deeper layers [62]. A residual block consists of two or three sequential convolutional layers and a shortcut (or skip) connection, which connects the input of the first layer with the output of the last [63]. This network has only one fully connected layer for performing the classification. The ResNet can have different depths, such as 34, 50, 101, or 152 layers [53].

The VGGNet [52] is still one of the most popular image recognition architectures serving as the basis of ground-breaking object recognition models. The main contribution of the VGG project was to demonstrate that the depth of a network is a critical component in CNN for obtaining good results in recognition or classification [64].

The Vision Transformer (ViT) [55] is a deep learning model based on the Transformer architecture, originally designed for natural language processing and later adapted for computer vision tasks. It segments images into small patches, treating them as tokens in a sequence, enabling the model to capture global relationships between different parts of an image. Leveraging the attention mechanism, ViT efficiently learns visual representations by highlighting relevant patterns at various scales. In addition to being highly competitive in image classification tasks, ViT can serve as a backbone for feature extraction in various computer vision applications, offering flexibility and high performance in large-scale scenarios [65].

ResNet was used in [32] to recognize morphological characteristics of herbarium specimens; it was also one of the models used by the four best-performing teams in the Herbarium 2019 competition for classifying herbarium plants of the Melastomataceae family [2]. VGG16 has also been used in studies with specimens from herbaria in [11, 39]. As far as we know, ViT has never been used to extract features from herbarium species but has shown competitive results in various computer vision applications [65, 66]

4.3 Classification

Four well-known classifiers were used to classify Piperaceae specimens: DT, *k*-NN, MLP, and SVM. The hyperparameters were optimized through a grid search [67]. The obtained values are shown in Table 6.

4.4 Evaluation metrics

The segmentation and classification experiments employed the cross-validation technique (5-fold). At the end of each execution, we computed the appropriate metric, and after the five executions, we calculated their respective average and standard deviation.

The Sørensen-Dice Coefficient [68, 69] was used to evaluate the image segmentation. It assesses the similarity between two regions through their spatial overlap [70]. The Sørensen-Dice Coefficient *SDC* is computed by (1):

$$SDC = \frac{2 * |a \cap b|}{|a| + |b|}, \tag{1}$$

where a is predicted area and b is ground truth [71]. The *SDC* ranges from 0 to 1, where $SDC = 1$ represents that the predicted segmentation area equals the ground truth.

F1-Score and Top-k accuracy are used to evaluate classification experiments. F1-Score consists of the harmonic mean between precision and recall as computed by (2).

$$F1\text{- }Score = 2 * \frac{precision * recall}{precision + recall}, \tag{2}$$

where *precision* measures the fraction of true positives in the set of identified positives, and *recall* measures the fraction of true positives identified among all the positives in the dataset [72]. The highest possible value of an F1-score is 1, showing perfect precision and recall, while the lowest possible value is 0 when either precision or recall is zero.

Top-k accuracy is a metric for evaluating multi-class classifiers, which counts the frequency of the true class among the highest-ranked predicted classes (Top-k). If $\hat{f}_{i,j}$ is the predicted class of the ith sample corresponding to the jth highest predicted score, and y_i is the corresponding ground truth, then the fraction of correct predictions on $n_{samples}$ is defined by (3) [73].

Table 6 Main hyperparameters used in classifiers

Classifier	Hyperparameter	Value
DT	max_depth	10
k-NN	n_neighbors	10
	weights	distance
MLP	activation	logistic
	learning_rate_init	0.01
	momentum	0.4
SVM	kernel	*rbf*

$$Top\text{-}\ k\ accuracy\ (y, \hat{f}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \sum_{j=1}^{k} 1(\hat{f}_{i,j} = y_i). \quad (3)$$

When $k = 1$, we have conventional accuracy. Typical values of k are 3, 5, or 10, but the appropriate value is context-dependent.

5 Experiments and discussion

Our experiments are divided into two distinct parts. The first part of our work addresses the segmentation, while the second is dedicated to tackling the classification problem. The hyperparameters were chosen using 5-fold cross-validation. Segmentation experiments were evaluated using the mean Sørensen-Dice coefficient. Mean F1-Score and Top-k describe classification experiments. The standard deviation was calculated for all experiments.

The datasets employed in segmentation and classification experiments are different. In the first case, the same dataset proposed by Kajihara et al. [29] was used. It includes 375 images of the Piperaceae family. Classification experiments deal with datasets from the state of Paraná, five Brazilian regions, and Brazil, as described in detail in Section 3.

5.1 Herbarium specimen segmentation

Initially, segmentation experiments were performed using U-Net. Two aspects of images were evaluated: dimensions (256×256, 400×400, and 512×512 pixels) and color modes (grayscale and RGB). We used 75% of the dataset for training, 5% for validation, and the 20% reminding to test. The 5-fold cross-validation was applied during all experiments and the mean Sørensen-Dice coefficient is described in Table 7. The results indicated that different color modes and image dimensions did not affect the task. Figure 9 shows two exsiccatae and their versions of segmentation: the first using a mask produced by a human and another using a predicted mask by U-Net. The segmentation of *Piper umbellata* was successful, while the segmentation of *Piper aduncum* failed. Failures usually occur when the background color resembles the dried plant.

To assess the impact of U-Net segmentation on the subsequent classification task, we applied the same protocol proposed by Kajihara et al. [29] but now using U-Net segmented images. We employed the same database, descriptors, and classifiers. The results of both studies were similar, indicating that the utilization of U-Net did not impact the species classification significantly. The accuracy using U-Net was 80%, while the accuracy for the original experiment was 80.53%.

Having established the feasibility of automated segmentation, we applied the U-Net for the database of Piperaceae described in Section 3. As image parameters did not significantly affect segmentation, we generated six dataset versions combining different image dimen-

Table 7 Mean Sørensen-Dice coefficients (%) in image segmentation approach

Dimension (pixels)	Color mode	
	RGB	Grayscale
256×256	98.10	97.72
400×400	98.16	97.81
512×512	97.95	97.57

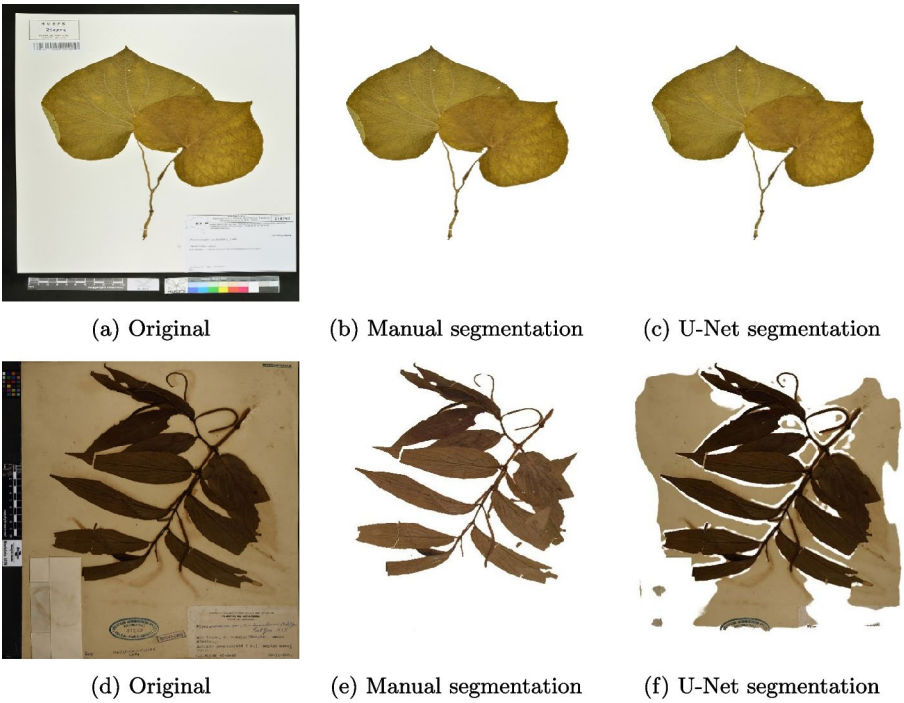


Fig. 9 Example of two exsiccatae: (**a**) *Piper umbellata* original sample [8], (**b**) segmented by human (**c**) and segmented by U-Net, (**d**) *Piper aduncum* original sample [8], (**e**) segmented by human (**f**) and segmented by U-Net

sions with color mode. This approach allowed us to assess the impact of these parameters on the classification process.

The following section describes the results using different datasets, descriptors, and classifiers used in classifying species of the Piperaceae family.

5.2 Herbarium specimen classification

The classification process is outlined in two sections. The first section details the evaluation of feature descriptors, classifiers, and image parameters. For these experiments, we use the subset of the state of Paraná. This subset was chosen because it has the highest average rate of images per species in Brazil. The bold entries in tables refer to the best results for each experiment. The research institution involved in this study is located in this state. The dataset used is described in detail in Table 1. In the second section, we utilized the most effective feature descriptors, classifiers, and parameters identified in the previous section to evaluate five regions besides the entire Brazilian territory. Details about the datasets used can be found in Table 2 and Table 3.

Table 8 Mean F1-Score using images 256×256 pixels - Paraná State Dataset (55 species and at least five images per class)

Classifier	Feature Descriptor					
	LBP	SURF	MobileNetV2	ResNet50	VGG16	ViT
DT	0.13	0.20	0.16	0.16	0.17	0.13
k-NN	0.14	0.24	0.23	0.19	0.20	0.29
MLP	0.19	0.30	0.35	0.35	0.35	**0.45**
SVM	0.11	0.25	0.31	0.32	0.29	0.37

Table 9 Mean F1-Score using images 400×400 pixels - Paraná State Dataset (55 species and at least five images per class)

Classifier	Feature Descriptor					
	LBP	SURF	MobileNetV2	ResNet50	VGG16	ViT
DT	0.14	0.18	0.18	0.14	0.18	0.14
k-NN	0.14	0.24	0.28	0.21	0.28	0.31
MLP	0.24	0.33	0.41	0.43	0.44	**0.48**
SVM	0.11	0.27	0.38	0.39	0.40	0.41

5.2.1 Evaluation of image parameters, descriptors, and classifiers

The dataset used in these experiments is unbalanced and contains 55 classes with at least five samples per class. The maximum number of samples was unrestricted. The classifier hyperparameters were optimized through grid-search with five folds. The mean F1-Score of five executions was used to evaluate the performance of experiments.

The first experiment evaluates the robustness of the feature descriptors and classifiers used. We evaluated four classifiers and six feature extraction methods, including two handcrafted (LBP and SURF) and four based on deep features or representation learning (MobileNetV2, ResNet50, VGG16 and ViT). In this experiment, images were grayscale and had a dimension of 256×256 pixels. We can note in Table 8 that the MLP classifier combined with ViT showed the best performance. The non-handcrafted features also showed superior results to the handcrafted ones. Contour features extracted by pre-trained models showed better performance than texture features. Experiments with LBP excluding white/white transitions to avoid computing background information from the image were conducted, but performance was not improved. Upon detailed evaluation of MLP and SVM classifiers, we observed that in cases with few samples, both made mistakes in similar situations.

We conducted a second experiment to explore the impact of varying image dimensions. The results for dimensions 400×400 and 512×512 pixels are described in Tables 9 and 10, respectively. In most cases, they indicate that larger images yield better rates of F1-Score. This may have occurred because details are more perceptible in images with larger dimensions, and some extracted features are probably essential to differentiate one species from another. Once again, MLP trained on features extracted from ViT delivered the best performance.

A third experiment evaluated the impact of the color modes (grayscale and RGB). Only non-handcrafted feature extractors were evaluated in the RGB mode since LBP [56] and SURF [59] are designed for grayscale images. In addition, the decision to focus on non-handcrafted methods was supported by their superior performance in the previous experi-

Table 10 Mean F1-Score using images 512×512 pixels - Paraná State Dataset (55 species and at least five images per class)

Classifier	Feature Descriptor					
	LBP	SURF	MobileNetV2	ResNet50	VGG16	ViT
DT	0.12	0.21	0.15	0.19	0.21	0.13
k-NN	0.14	0.26	0.32	0.28	0.31	0.31
MLP	0.25	0.35	0.45	0.47	0.48	**0.50**
SVM	0.10	0.26	0.42	0.45	0.44	0.42

Table 11 Mean F1-Score using RGB images with 512×512 pixels - Paraná State Dataset (55 species and at least five images per class)

Classifier	Feature Descriptor			
	MobileNet-V2	ResNet50	VGG16	ViT
DT	0.18	0.16	0.21	0.14
k-NN	0.32	0.26	0.28	0.31
MLP	0.47	0.48	**0.51**	**0.51**
SVM	0.42	0.45	0.44	0.45

Table 12 Mean F1-Score using RGB images with 512×512 pixels - Paraná State Dataset (36 species and at least 10 images per class)

Classifier	Feature Descriptor			
	MobileNetV2	ResNet50	VGG16	ViT
DT	0.22	0.23	0.23	0.16
k-NN	0.36	0.30	0.34	0.37
MLP	0.53	0.53	**0.56**	**0.56**
SVM	0.48	0.49	0.50	0.51

Table 13 Mean F1-Score using RGB images with 512×512 pixels - Paraná State Dataset (23 species and at least 20 images per class)

Classifier	Feature Descriptor			
	MobileNetV2	ResNet50	VGG16	ViT
DT	0.21	0.28	0.30	0.23
k-NN	0.45	0.35	0.41	0.44
MLP	0.63	0.60	**0.64**	0.63
SVM	0.56	0.58	0.60	0.59

ment. The results are described in Table 11. The most classifiers achieved high rates when using RGB images. The highest F1-Score achieved was also obtained by MLP classifier but now using features extracted from VGG16 and ViT.

Despite efforts, the F1-Score remains below 0.51 in the best cases. Thus, we decided to evaluate the impact by increasing the minimum number of samples per species. Sets with a minimum of 10 and 20 images of each species were evaluated. As seen in Tables 12 and 13, the results showed that increasing the number of samples per species leads to improved model performance. On the other hand, the number of classes was significantly reduced since we used only classes with at least 10 and 20 samples. The number of classes using a minimum of 10 samples per species was 36, and using 20 examples per species was 23.

The best results were consistently achieved using MLP across all experiments. In most cases, ViT was utilized as a feature extractor for these top-performing results. We can also

notice that VGG16 becomes more competitive in subsets with more restrictive sample numbers.

The best result using a minimum of five images per species was 0.51 (±0.02), whereas a subset with a minimum of 20 images per species yielded a higher F1-Score of 0.64 (±0.03). The increase of 0.13 percentage points (in the best case) from five to twenty samples per class may have been influenced by the decrease in the number of classes. There was a reduction of about 41.82% in the number of classes and, at the same time, an improvement of 0.13 percentage points. Therefore, it is difficult to say that increasing the number of samples can improve the performance of species classification.

Finally, we evaluate performance using Top-k, with $k = \{3, 5\}$. The main goal of Top-k is to collaborate with the taxonomist to show a hit list where there is a greater possibility of the sample seeking to be included. Table 14 describes the results achieved for the Paraná state subset using a minimum number of images 5, 10, and 20 per species. MLP was trained using features extracted from VGG16 and ViT. As expected, the performance improved when we used $k = 3$ and $k = 5$.

Building a database with many samples of each species is challenging because, in herbarium collections, some species are represented by few specimens, while others are represented by many [2]. Several factors contribute to this situation, such as the process of collecting samples, as some species are more abundant in nature than others [36]; difficult accessing certain collection sites [25]; and in some species, a particular plant organ, such as the fruit, is essential for the accurate identification of the species but may not be available year-round or is not easily collected [28].

In short, our experiments with the Paraná State dataset demonstrated that the most successful approach for classifying herbarium specimens involved the utilization of an MLP classifier in conjunction with features extracted from VGG16 and ViT. This approach utilized RGB images of 512×512 pixels for each Piperaceae species. The following section reports the results using this set of features, classifiers, and sizes for other subsets.

5.3 Evaluating five regions from Brazil

This section employed the parameters from experiments on the Paraná State dataset. All the following experiments used the MLP classifier and features extracted from VGG16 and ViT. We also used RGB images and dimensions of 512×512 pixels. The results of these experiments are described in Table 15.

In Table 15, the experiments indicated that, in most cases, the F1-Score means achieved using ViT (12 of 15 tests) was better than VGG16 (two of 15 tests), and in only one situation, the rate was similar. The variation between the results of the two descriptors was 1% to 4%.

We also performed experiments with all species found in the Brazilian territory using features extracted from the VGG16 and ViT with the MLP classifier. The F1-Score is shown in Table 16. Similar to the previous experiments, the more restrictive subset yields bet-

Table 14 Top-k (%) using RGB images with 512×512 pixels - Paraná State Dataset

# Minimum samples	Species/classes	VGG16		ViT	
		Top-3	Top-5	Top-3	Top-5
5	55	77.77	85.46	78.36	88.18
10	36	82.35	89.55	82.51	91.09
20	23	86.67	93.33	85.82	92.58

Table 15 Mean F1-Score using RGB images with 512×512 pixels - Regions of Brazil Dataset

Regional subsets	# Minimum samples	Species/classes	VGG16	ViT
North	5	107	0.32	0.34
	10	68	0.37	0.40
	20	41	0.46	0.50
Northeast	5	48	0.49	0.52
	10	35	0.57	0.61
	20	21	0.65	0.66
Midwest	5	42	0.55	0.56
	10	29	0.60	0.61
	20	17	0.68	0.66
South	5	72	0.48	0.51
	10	49	0.54	0.56
	20	33	0.59	0.58
Southeast	5	102	0.39	0.40
	10	67	0.43	0.46
	20	42	0.51	0.51

Table 16 Mean F1-Score, using the combination of MLP with VGG16 and ViT on 512× 512 pixels RGB images - Brazil Dataset

# Minimum samples	Species/classes	VGG16	ViT
5	236	0.38	0.37
10	160	0.41	0.40
20	106	0.41	0.45

ter performance, but almost all (two of three tests) the cases from features extracted were VGG16.

We also evaluated the performance using the Top-k metric for the five regions of Brazil. The results in Table 17 show that the ViT (20 of 30 tests) is better than VGG16 (10 of 30 tests), and even with a large number of classes (minimum number of samples equals five), it is possible to achieve rates between 57.90% and 79.37% using the Top-3. When the number of classes is more restrictive, the Top-3 rate vary from 72.10% to 88.91%.

Finally, we calculate the Top-k for the experiments conducted on the Brazil dataset, as shown in Table 18. Here, ViT achieved a better performance than VGG16 for all subsets. Taking into account the Top-3 and Top-5 metrics, we can conclude that the results obtained are not only encouraging but also competitive with the leading works in the literature, particularly when considering a botanical family with intricate identification challenges. We achieved 63.60% and 72.17% performance, employing 236 classes (species) to Top-3 and Top-5, respectively. These rates are particularly remarkable considering the challenges posed by significant interclass similarity and intraclass variation.

We explore various diverse subsets, handcrafted and non-handcrafted features, and different performance metrics. The performance of these experiments establishes a baseline for future research.

5.4 Discussion

Upon analyzing the confusion matrix generated and inspecting the species images, we observed species with many errors. These species are usually very similar, that is, high

Table 17 Top-k (%), using the combination of MLP with VGG16 and ViT on 512×512 pixels RGB images - Regions of Brazil Dataset

Regional subsets	# Minimum sample	Species/classes	VGG16		ViT	
			Top-3	Top-5	Top-3	Top-5
North	5	107	56.53	64.54	57.90	65.99
	10	68	62.99	72.10	64.65	73.62
	20	41	69.98	78.71	72.10	80.29
Northeast	5	48	79.67	88.14	79.37	88.41
	10	35	85.83	92.71	85.21	91.48
	20	21	89.56	94.46	88.91	93.30
Midwest	5	42	80.63	87.68	79.31	87.86
	10	29	82.95	90.13	83.04	90.13
	20	17	88.18	94.37	88.06	94.14
South	5	72	75.78	83.88	78.63	87.01
	10	49	80.60	88.61	82.02	91.47
	20	33	83.59	91.51	85.59	92.94
Southeast	5	102	66.65	76.23	65.46	75.30
	10	67	71.08	81.14	72.87	82.49
	20	42	76.98	86.16	79.38	88.20

Table 18 Top-k (%), using the combination of MLP with VGG16 and with ViT on 512× 512 pixels RGB images - Brazil Dataset

# Minimum samples	Species/classes	VGG16		ViT	
		Top-3	Top-5	Top-3	Top-5
5	236	61.18	69.31	63.60	72.17
10	160	63.83	72.43	67.24	75.82
20	106	69.01	78.05	71.32	80.60

interclass similarity. In practice, this is common, and often, experts need more information to label them correctly. Information such as plant location, collection time, fruits, and other details can be considered to increase accuracy in the classification step.

Our models could not correctly identify any samples of the species *Peperomia blanda*. One of the test examples of this species was classified as *Peperomia glabella*, and the similarities between them can be observed in Fig. 10.

Another factor that may have impacted the predictions is the intraclass variability. For example, the *Piper abutiloides* has four samples as illustrated in Fig. 11.

The experiments revealed that the MLP classifier with non-handcraft features, especially ViT or VGG16, performed better for classifying Piperaceae herbarium specimens. In most tests evaluated, ViT achieved a rate more than VGG16. The difference varies between 0.02 and 0.10 for the feature descriptors. Although ViT achieved a better accuracy, VGG16 is a promising computational alternative for this work, as it delivers similar results while using only half the features of ViT.

In summary, the studies presented in this work investigate the Piperaceae family with many species. Table 19 shows the performance of the leading studies in the field to date. Despite the lack of standardized metrics, it highlights the challenges posed by our dataset compared to others. The different datasets used in this work are available so that new research can be carried out.

(a) *Peperomia blanda* (b) *Pepeormia glabella*

Fig. 10 Similarity between two species: (**a**) *Peperomia blanda* (**b**) *Peperomia glabella*. Original exsiccatae from [8]

Fig. 11 Example of intraclass variation of the species *Piper abutiloides*. Original exsiccatae from [8]

6 Conclusion

Automated species identification holds significant potential as a valuable tool for taxonomists and technical staff, streamlining identification and bringing possible misidentifications of herbarium specimens to the attention of the responsible curator. However, the main contributions rely on diverse taxonomic groups aggregated on large datasets, diverging from herbaria's practical realities. To overcome this challenge, we presented a curated dataset of herbarium specimens of the Piperaceae botanical family with plants collected in several Brazilian regions, using the *species*Link repository and collaborating with domain experts. Our dataset has 236 species and 10,503 digitized images of exsiccatae from 38 herbaria.

We accompany the dataset with a robust experimental protocol for segmentation, feature extraction, and classification, with results that establish a baseline for future research on species identification using digitized herbarium images. The baseline underscores the superior performance of non-handcrafted features compared to handcrafted ones. The rates achieved are closely related to the complexity of identifying the Piperaceae family with interclass similarity and intraclass differences, as experts have pointed out. The high imbalance degree may also have contributed to the low F1-Score rates in species classification.

Table 19 Summary of state of the art on herbarium specimen identification

Contribution	Dataset	# Classes	Performance (%)
Wijesingha and Marikar [25]	Private	17	$85.0^{1,a}$
Clark et al. [24]	Private	4	$44.0^{1,a}$
Unger et al. [27]	Public	17	$84.88^{1,a}$
Wilf et al. [30]	Public	1419	$72.14^{2,c}$ 57.26 ^{2,d}
Grim et al. [26]	Private	6	94 to $100^{1,a}$
Kho et al. [28]	Private	3	$83.30^{1,a}$
Carranza-Rojas et al. [12]	Public	1,204 255	$70.3^{1,a}$ 79.6 ^{1,a}
Schuettpelz et al.[31]	Private	2	$96^{1,d}$
Younis et al. [32]	Private	1,000	$82.40^{1,a}$
Carranza-Rojas et al. [36]	Public	1,191 498 124	$64.32^{1,a}$ 76.23 ^{1,b} $88.17^{1,d}$
Pryer et al. [11]	Public	3	90.0 ^{1,a}
Little et al. [2]	Public	683	88.0 ^{1,a}
Shirai et al. [35]	Private	2171	$96.4^{1,a}$
Kajihara et al. [29]	Public	5	$80.53^{1,b}$
Pravin and Deepa [45]	Private	15	$88.0^{1,a}$
Lutio et al. [37]	Public	64,500	$84.5^{1,a}$
Our - Brazil	Public	236	71.32 ^{3,a}
Our - Paraná	Public	55	78.36 ^{3,a}

1. Accuracy; 2. Mean Accuracy; 3. Top-3; a. Species; b. Genus; c. Order; d. Family

On the other hand, the Top-k rates encourage the development of automatic identification systems, supporting the specialists and users of herbaria.

In future work, we plan to evaluate additional feature extractors and explore whether data balancing techniques can enhance the performance of models in identifying herbarium specimens. In addition, we intend to evaluate other botanical families to determine whether the proposed approach performs similarly across different families. Another possibility is to expand the dataset by including samples from other countries, enabling the creation of a larger database and allowing us to assess performance in a cross-dataset scenario.

Author contributions A. Y. Kajihara: Study conception, Data acquisition, Implementation, and Writing - original draft. G. A. Queiroz: Data acquisition and Data curation. M. G. Caxambu: Data acquisition and Data curation. L. E. S. Oliveira: Supervision and Critical revision. D. Bertolini: Algorithm design, Results interpretation, Critical revision, and Writing - review and editing. A. L. Schwerz: Supervision, Results interpretation, Critical revision, and Writing - review and editing.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and the Federal University of Technology - Paraná.

Data availability The dataset and code are available at https://doi.org/10.5281/zenodo.14599766.

Declarations

Ethics approval Not applicable.

Competing interests The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Thiers BM (2022) The world's herbaria 2021: A summary report based on data from index herbariorum, New York. https://sweetgum.nybg.org/science/wp-content/uploads/2022/02/The_Worlds_Herbaria_Jan_2022.pdf Accessed 01 March 2024
2. Little DP, Tulig M, Tan KC, Liu Y, Belongie S, Kaeser-Chen C, Michelangeli FA, Panesar K, Guha R, Ambrose BA (2020) An algorithm competition for automatic species identification from herbarium specimens. Appl Plant Sci 8(6):11365. https://doi.org/10.1002/aps3.11365
3. Thiers BM (2020) Herbarium: The Quest to Preserve and Classify the World's Plants. Timber Press, Portland
4. Ebach M, Valdecasas AG, Wheeler Q (2011) Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. Cladistics 27(5):550–557. https://doi.org/10.1111/j.1096-0031.2011.00348.x
5. GBIF (2022) Global Biodiversity Information Facility. https://www.gbif.org/what-is-gbif Accessed 01 March 2024
6. iDigBio (2021) Integrated Digitized Biocollections. https://www.idigbio.org/portal Accessed 01 March 2024
7. Reflora (2020) Herbário Virtual. http://reflora.jbrj.gov.br/reflora/herbarioVirtual Accessed 01 March 2024
8. speciesLink (2024) speciesLink Network. https://specieslink.net/ Accessed 01 March 2024
9. CRIA (2018) Centro de Referência em Informação Ambiental. http://www.cria.org.br/projetos Accessed 01 March 2024
10. Peixoto AL, Maia LC (2013) Manual de Procedimentos Para Herbários. Editora Universitária UFPE, Recife
11. Pryer K, Tomasi C, Wang X, Meineke E, Windham M (2020) Using computer vision on herbarium specimen images to discriminate among closely related horsetails (Equisetum). Aplic Plant Sci 8(6):11369. https://doi.org/10.1002/aps3.11372
12. Carranza-Rojas J, Goëau H, Bonnet P, Mata-Montero E, Joly A (2017) Going deeper in the automated identification of herbarium specimens. BMC Evol Biol 17(181):2–14. https://doi.org/10.1186/s12862-017-1014-z
13. Mata-Montero E, Carranza-Rojas J (2016) Automated plant species identification: Challenges and opportunities. In: Proc. 6th IFIP world information technology forum (WITFOR'16), San José, Costa Rica, pp 26–36. https://doi.org/10.1007/978-3-319-44447-5_3
14. Stroud S, Fennell M, Mitchley J, Lydon S, Peacock J, Bacon KL (2022) The botanical education extinction and the fall of plant awareness. Ecol Evol 12(7):9019. https://doi.org/10.1002/ece3.9019
15. Godfray HCJ (2002) Challenges for taxonomy. Nature 417(6884):17–19. https://doi.org/10.1038/417017a
16. Joly A, Goëau H, Kahl S, Deneu B, Servajean M, Cole E, Picek L, Ruiz de Castañeda R, Bolon I, Durso A, Lorieul T, Botella C, Glotin H, Champ J, Eggel I, Vellinga W-P, Bonnet P, Müller H (2020) Overview of LifeCLEF 2020: A system-oriented evaluation of automated species identification and species distribution prediction. In: Proc. 11th Int. Conf. the CLEF Association (CLEF'20), Thessaloniki, Greece. https://doi.org/10.1007/978-3-030-58219-7_23
17. Mai P, Rossado A, Bonifacino JM, Waechter JL (2016) Taxonomic revision of *Peperomia* (Piperaceae) from Uruguay. Phytotaxa 244(2):125–144. https://doi.org/10.11646/phytotaxa.244.2.2
18. Mathieu G, Callejas R (2006) New synonymies in the genus *Peperomia* Ruiz & Pav. (Piperaceae): An annotated checklist. Candollea 61(2):331–363
19. Jaramillo MA, Rodríguez-Duque D, Escobar-Alba M (2023) A new species of Piper (Piperaceae) with peltate leaves from Serranía de las Quinchas. Colombia PhytoKeys 227:9–24. https://doi.org/10.3897/phytokeys.227.101405
20. Callejas-Posada R (2020) Piperaceae. In: Davidse G, Ulloa CU, Arbeláez AL, Hernández H, Knapp S, Bilsborrow T, Palacio M, Gunter D (eds) Flora Mesoamericana: Piperaceae, vol 2. Universidad Nacional Autónoma de México, México, pp 1–590
21. Christ JA, Sarnaglia-Junior VB, Barreto LM, Guimarães EF, Garbin ML, Carrijo TT (2016) The genus Piper (Piperaceae) in the Mata das Flores State Park, Espírito Santo. Brazil. Rodriguésia 67(4):1031–1046. https://doi.org/10.1590/2175-7860201667413
22. Chulif S, Lee SH, Chang YL, Chai KC (2022) A machine learning approach for cross-domain plant identification using herbarium specimens. Neural Comput Appl 35(8):5963–5985. https://doi.org/10.1007/s00521-022-07951-6

23. Hussein BR, Malik OA, Ong W, Slik JWF (2022) Application of computer vision and machine learning for digitized herbarium specimens: A systematic literature review. Ecol Inform 69. https://doi.org/10.1016/j.ecoinf.2022.101641
24. Clark J, Corney D, Tang H (2012) Automated plant identification using artificial neural networks. In: Proc. IEEE Symp. Comput. Intell. Bioinf. Comput. Biol. (CIBCB'12) San Diego, CA, USA, pp 343–348. https://doi.org/10.1109/CIBCB.2012.6217250
25. Wijesinghe D, Marikar F (2012) Automatic detection system for the identification of plants using herbarium specimen images. Tropical Agric Res 23(1):42–50. https://doi.org/10.4038/tar.v23i1.4630
26. Grimm J, Hoffmann M, Stöver B, Müller K, Steinhage V (2016) Image-based identification of plant species using a model-free approach and active learning. In: Proc. 39th Annual German Conference on AI, Klagenfurt, Austria, pp 169–176. https://doi.org/10.1007/978-3-319-46073-4_16
27. Unger J, Merhof D, Renner S (2016) Computer vision applied to herbarium specimens of german trees: Testing the future utility of the millions of herbarium specimen images for automated identification. BCM Evol Biol 16(248) .https://doi.org/10.1186/s12862-016-0827-5
28. Kho S, Manickam S, Malek S, Mosleh M, Dhillon SK (2017) Automated plant identification using artificial neural network and support vector machine. Front Life Sci 10(1):98–107. https://doi.org/10.1080/21553769.2017.1412361
29. Kajihara AY, Bertolini D, Schwerz AL (2022) Identification of herbarium specimens: A case study with Piperaceae Giseke family. In: Proc. 29th Int. Conf. Syst., Signals Image Process. (IWSSIP'22), pp 1–4 (2022). https://doi.org/10.1109/IWSSIP55020.2022.9854444
30. Wilf P, Zhang S, Chikerrur S, Little S, Wing S, Serre T (2016) Computer vision cracks the leaf code. Proc Natl Acad Sci USA 113(12):3305–3310. https://doi.org/10.1073/pnas.1524473113
31. Schuettpelz E, Frandsen P, Dikow R, Brown A, Orli S, Peters M, Metallo A, Funk V, Dorr L (2017) Applications of deep convolutional neural networks to digitized natural history collections. Biodivers Data J 5:21139. https://doi.org/10.3897/BDJ.5.e21139
32. Younis S, Weiland C, Hoehndorf R, Dressler S, Hickler T, Seeger B, Schmidt M (2018) Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. Bot Lett 165(3–4):377–383. https://doi.org/10.1080/23818107.2018.1446357
33. Tan KC, Liu Y, Ambrose B, Tulig M, Belongie S (2019) The Herbarium Challenge 2019 Dataset. arXiv:1906.05372 Accessed 01 March 2024
34. de Lutio R, Little D, Ambrose B, Belongie SJ (2021) The Herbarium 2021 Half-Earth Challenge Dataset. arXiv:2105.13808
35. 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. Sci Rep 12(8066). https://doi.org/10.1038/s41598-022-11450-y
36. Carranza-Rojas J, Joly A, Goëau H, Mata-Montero E, Bonnet P (2018) Automated identification of herbarium specimens at different taxonomic levels. In: Joly, A., Vrochidis, S., Karatzas, K., Karppinen, A., Bonnet, P. (eds.) Multimedia Tools and Applications for Environmental & Biodiversity Informatics, pp 151–167. Springer, Cham . https://doi.org/10.1007/978-3-319-76445-0_9
37. de Lutio R, Park JY, Watson KA, D'Aronco1, 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. Front Plant Sci 12. https://doi.org/10.3389/fpls.2021.787127
38. White A, Dikow R, Baugh M, Jenkins A, Frandsen P (2020) Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning. Appl Plant Sci 8(6):11352. https://doi.org/10.1002/aps3.11352
39. Triki A, Bouaziz B, Mahdi W, Hamed H, Gaikwad J (2022) Deep learning based approach for digitized herbarium specimen segmentation. Multimed Tools Appl 81:28689–28707. https://doi.org/10.1007/s11042-022-12935-8
40. Milleville K, Chandrasekar KKT, Weghe NV, Verstockt S (2023) Evaluating segmentation approaches on digitized herbarium specimens. In: 18th Int. Symp. on Vis. Comput. (ISVC'23) Lake Tahoe, NV, USA, pp 65–78. https://doi.org/10.1007/978-3-031-47966-3_6
41. Milleville K, Chandrasekar KKT, Verstockt S (2023) Automatic extraction of specimens from multi-specimen herbaria. ACM J Comput Cult Herit 16(1):4. https://doi.org/10.1145/3575862
42. Younis S, Schmidt M, Weiland C, Dressler S, Seeger B, Hickler T (2020) Detection and annotation of plant organs from digitised herbarium scans using deep learning. Biodivers Data J 8:57090. https://doi.org/10.3897/BDJ.8.e57090
43. Triki A, Bouaziz B, Mahdi W, Gaikwad J (2020) Objects detection from digitized herbarium specimen based on improved YOLO V3. In: Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Applic. (VISAPP'20) vol 4. Valletta, Malta, pp 523–529. https://doi.org/10.5220/0009170005230529

44. Thompson KM, Turnbull R, Fitzgerald E, Birch JL (2023) Identification of herbarium specimen sheet components from high-resolution images using deep learning. Ecol Evol 13(8):10395. https://doi.org/10.1002/ece3.10395
45. Pravin A, Deepa C (2022) Piper plant classification using deep CNN feature extraction and hyperparameter tuned Random Forest classification. Transdiscipl J Eng Sci 13:233–258. https://doi.org/10.22545/2022/00202
46. Hussein BR, Malik OA, Ong W-H, Slik JWF (2019) Semantic segmentation of herbarium specimens using deep learning techniques. In: Proc. 6th Int. Conf. Computational Science and Technology (ICCST'19), Kota Kinabalu, Malaysia, pp 321–330. https://doi.org/10.1007/978-981-15-0058-9_31
47. Ondov BD, Bergman NH, Phillippy AM (2013) Krona: Interactive metagenomic visualization in a web browser. In: Nelson, K.E. (ed.) Encyclopedia of Metagenomics, pp 1–8. Springer, New York.https://doi.org/10.1007/978-1-4614-6418-1_802-1
48. Ortigosa-Hernández J, Inza I, Lozano JA (2017) Measuring the class-imbalance extent of multi-class problems. Pattern Recognit Lett 98(15):32–38. https://doi.org/10.1016/j.patrec.2017.08.002
49. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI'15), Munich, Germany, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
50. Yin X-X, Sun L, Fu Y, Lu R, Zhang Y (2022) U-Net-based medical image segmentation. J Healthc Eng 2022. https://doi.org/10.1155/2022/4189781
51. Hussein BR, Malik OA, Ong W-H, Slik JWF (2021) Reconstruction of damaged herbarium leaves using deep learning techniques for improving classification accuracy. Ecol Inform 61. https://doi.org/10.1016/j.ecoinf.2021.101243
52. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: 3rd Int. Conf. Learn. Represent. (ICLR'15), San Diego, CA, USA, pp 1–14. https://doi.org/10.48550/arXiv.1409.1556
53. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR'16), Las Vegas, NV, USA, pp 770–778 . https://doi.org/10.1109/CVPR.2016.90
54. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. https://doi.org/10.48550/arXiv.1704.04861
55. 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. arXiv:2010.11929 Accessed 01 Janery 2025
56. Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 29(1):51–59. https://doi.org/10.1016/0031-3203(95)00067-4
57. Liu L, Fieguth PW, Guo Y, Wang X, Pietikäinen M (2017) Local binary features for texture classification: Taxonomy and experimental study. Pattern Recognit 62:135–160. https://doi.org/10.1016/j.patcog.2016.08.032
58. Vukicevic A, Zabotti A, de Vita S, Filipovic N (2018) Assessment of machine learning algorithms for the purpose of primary Sjögren's syndrome grade classification from segmented ultrasonography images. In: Proc 3rd Int. Conf. Future Access Enablers Ubiquitous Intell. Infrastructures (FABULOUS'17), Bucharest, Romania, pp 239–245. https://doi.org/10.1007/978-3-319-92213-3_35
59. Bay H, Tuytelaars T, Gool LV (2006) SURF: Speeded Up Robust Features. In: Proc. 9th european conference on computer vision (ECCV'06), Graz, Austria, pp 404–417.https://doi.org/10.1007/11744023_32
60. Hirasen D, Viriri S (2020) Plant species recognition using Local Binary and Local Directional Patterns. In: Proc. 2nd Int. Multidiscipl. Inf. Technol Eng. Conf. (IMITEC'20), Kimberley, South Africa, pp 1–9 (2020). https://doi.org/10.1109/IMITEC50163.2020.9334091
61. George J, Gladston Raj S (2019) Leaf identification using harris corner detection, surf feature and flann matcher. International Journal of Innovative Technology and Exploring Engineering (IJITEE). 8(11):2136–2143. https://doi.org/10.35940/ijitee.K2016.0981119
62. Shanmugamani R (2018) Deep Learning for Computer Vision. Packt Publishing, Birmingham
63. Vasilev I (2019) Advanced Deep Learning with Python. Packt Publishing, Birmingham
64. Villán AF (2019) Mastering OpenCV 4 with Python. Packt Publishing, Birmingham
65. Touvron H, Cord M, Jégou H (2022) Deit iii: Revenge of the vit. In: Proc. 17th european conference on computer vision (ECCV'22), Tel Aviv, Israel, pp 516–533. https://doi.org/10.1007/978-3-031-20053-3_30
66. Pereira GML, Foleis JH, Souza Brito A, Bertolini D (2024) A database for soybean seed classification. In: Proc. 37th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'24), Manaus, Brazil, pp 1–6 .https://doi.org/10.1109/SIBGRAPI62404.2024.10716268

67. Scikit-Learn (2012) Model selection - GridSearchCV. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV Accessed 01 March 2024
68. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302. https://doi.org/10.2307/1932409
69. Sørenson TJ (1948) A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. I kommission hos E, Munksgaard, København
70. Veta M, van Diest PJ, Kornegoor R, Huisman A, Viergever MA, Pluim JPW (2013) Automatic nuclei segmentation in H &E stained breast cancer histopathology images. PLoS ONE 8(7):70221. https://doi.org/10.1371/journal.pone.0070221
71. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC (1994) Morphometric analysis of white matter lesions in MR images: Method and validation. IEEE Trans Med Imaging 4:716–724. https://doi.org/10.1109/42.363096
72. Lakshmanan V, Görner M, Gillard R (2021) Practical machine learning for computer vision: end-to-end machine learning for images. O'Reilly, Sebastopol
73. Scikit-Learn (2022) Metrics and scoring: quantifying the quality of predictions. https://scikit-learn.org/stable/modules/model_evaluation.html#top-k-accuracy-score Accessed 01 March 2024

Authors and Affiliations

Alexandre Yuji Kajihara[1] · George Azevedo de Queiroz[2] · Marcelo Galeazzi Caxambú[3] · Luiz Eduardo S. Oliveira[4] · Diego Bertolini[1] · André Luis Schwerz[1]

✉ Alexandre Yuji Kajihara
alexandrey@utfpr.edu.br

George Azevedo de Queiroz
georgeazevedo08@gmail.com

Marcelo Galeazzi Caxambú
mcaxambu@utfpr.edu.br

Luiz Eduardo S. Oliveira
luiz.oliveira@ufpr.br

Diego Bertolini
diegobertolini@utfpr.edu.br

André Luis Schwerz
andreluis@utfpr.edu.br

[1] Departamento Acadêmico de Computação, Universidade Tecnológica Federal do Paraná, Campo Mourão, Paraná, Brazil

[2] Departamento de Farmácia, Universidade do Estado do Rio de Janeiro, Campo Grande, Rio de Janeiro, Brazil

[3] Departamento Acadêmico de Biodiversidade e Conservação da Natureza, Universidade Tecnológica Federal do Paraná, Campo Mourão, Paraná, Brazil

[4] Departamento de Informática, Universidade Federal do Paraná, Curitiba, Paraná, Brazil