Politecnico di Torino

Politecnico di Torino

Data Science and Engineering

A.Y. 2024/2025

Graduation Session October 2025

XAI tools to predict biological invasiveness: a case study in plants

Relatori:

Prof. Daniele Apiletti

Dr. Simone Monaco

Candidati:

Guido Spina

Abstract

Invasive species are organisms that have been introduced (intentionally or accidentally) into an area where they are not originally present, and can have a negative impact on environment, economy or health by spreading quickly and without control. Their identification is important because it allows humans to either eradicate them if the spreading process has already begun, or to avoid their import into a new area altogether. To this moment, there is no method to identify what morphological traits make a species of plants potentially invasive and what makes it non-invasive based exclusively on image data, relying instead on categorical or numerical traits that are not always available. In this work we propose a pipeline to identify, within a family of plants, which species have the potential to be invasive and which ones have not, using the *Lythrum* genus as a case study. To do this we employ BioCLIP 2, a computer vision foundation model specialized in the biological domain, as a feature extractor to train a classifier to recognize an invasive species or a non-invasive one. Then, using Integrated Gradients as an explainability method, we highlight what are the regions of the image that the classifier identifies as most useful for its prediction. By extracting these regions and clustering them we are able to analyze what morphological traits are taken into consideration by the classifier, making them possible candidates as features that allow a species to be invasive. Additionally, it is possible to understand which traits or features drag the model into misclassification. With this work we are able to provide a pipeline to better explore and explain predictions on image data in the biological domain. For future works that will take on this problem it might be interesting to extend the reach of the study by taking into consideration other family of plants, and to integrate the outcome of the pipeline into already existing analysis of species invasiveness that utilize different features as predictors.

Acknowledgements

I would like to thank Dr. Simone Monaco for his constant supervision and support throughout the whole process of both experiments and writing. I would also like to thank Dr. Riccardo Ciarle for his help, supervision and advices on the biological side of this work, and for steering us onto the right direction when we were about to commit mistakes on biological topics. Finally thank you to Barbara Frittella for working on this thesis with me, being sharp and precise throughout the whole process.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Invasive species are organisms transported by humans, either intentionally or unintentionally, beyond their native range, where they establish, spread, and often cause ecological disruption.

Not all species are invasive: to be invasive a species needs to be 'alien' (i.e. non-native from a specific location), and a subgroup of alien species can be invasive.

Preventing the introduction of invasive species and managing their impacts is essential to prevent ecosystems from decline. Failure in doing so could create several issues to economy, food security and human health.

One of the most famous examples of disruption brought by an invasive species is the case of water hyacinth (*Pontederia crassipes*). It is a free-floating acquatic plant originary from South America, where it's one of the main food sources of the Amazonian manatee, which helps controlling its diffusion. When introduced in other ecosystems, however, water hyacinth has a huge impact: it is able to outcompete native acquatic plants, affecting their photosynthesis and their growth. By occupying all the available water surface, it inhibits photosynthesis in underwater plants, causing them to die off. As the dead plant material decomposes, this process consumes large amount of oxygen in the water, leading to oxygen depletion and ultimately causing fish kills.

Several health problems for humans can also be traced back to an excessive presence of water hyacinth. This plant is able to absorb large quantities of heavy metals and other substances toxic for humans: when it dies it rots and releases them, causing pollution and disrupting the quality of the water, in some cases even affecting the residents drinking water.

Very important are also the economical aspects of invasion from *Pontederia crassipes*: one of the many reasons for this is that heavy presence of water hyacinth in water bodies (e.g. rivers or lakes) can reduce or completely inhibit transports (either for humans or cargo). In the USA it has been given the nickname *'million dollars weed'*, not because of its value, but because of the significant sum of money

spent every year by local governments for its (often unsuccessful) removal attempts.

It is therefore important to be able to successfully recognize invasive species, either to mitigate the effects of the invasion process, or to prevent it from starting altogether.

There are several studies in the biological domain that investigate the potential invasiveness of a species, meaning the identification of traits that allow an alien species to prevail over both other alien species and native species, spreading when located in a new ecosystem. They demonstrated that predicting invasiveness using functional traits is possible, but they have important limitations: they are either limited in scale (either by working on a small set of species or in a tight geographical location) or they use a set of data which is not always easy to obtain (such as seed mass, dispersal mode and other similar numerical or categorical data), a type of information which might not be largely available for every species.

We are currently not aware of any research project that worked on the identification of an invasive species working exclusively with image data, and that is what we propose: we define a pipeline to predict possible invasiveness between species of the same genus through the exclusive use of visual traits.

We take as case study *Lythrum*, one of the genera of the *Lythraceae* family, an herbaceous (annual or perennial) genus. The *Lythrum* genus wast chosen since one of its species (*Lythrums salicaria*, commonly known as purple loosestrife, an herbaceous perennial plant native to Europe, Asia, northern Africa, and eastern Australia) has been inserted by IUCN in the '100 of the World's Worst Invasive Alien Species list'.

To reach our goal we gather images of the species belonging to the *Lythrum* genus from *iNaturalist*, a platform where a community of users can upload pictures of living organisms, labeling them and contributing to citizen science research projects. The good coverage of *Lythrum* on iNaturalist is one of the reasons for our choice of this genus: We were able to retrieve images for a total of 30 species, three of which are invasive (in locations where they are not native): *Lythrum hyssopifolia, Lythrum salicaria* and *Lythrum virgatum*.

One of the core tools for our work is BioCLIP, a state-of-the-art computer vision foundation model, pre-trained on biological data. It is built on OpenAI's CLIP framework, training jointly a vision encoder and a text encoder using contrastive learning with image-text pairs samples. BioCLIP has proved to be able to extract fine-grained representation of image data, detecting subtle biological structures and distinguishing between species with a similar appearence, or scarcely represented in the training dataset.

We use BioCLIP as a feature extractor to map images in the dataset into multi-dimensional embeddings, used to train a classifier able to distinguish invasive species from non-invasive.

Additionally we increase the interpretability of the pipeline by introducing

explainability: we generate attribution maps for each image using Integrated Gradients, an algorithm able to identify the regions that drive the model prediction the most.

We extract these regions and cluster them to detect what morphological structure they represent: these makes us able to identify what traits, characteristic for each species, are present in each image, linking them to the model prediction.

Finally we analyze this results to extract patterns and obtain information on the morphological visual traits that influence the potential of a species to be invasive or non-invasive.

This thesis is organized with the following structure:

- Chapter 2 reviews related literature, including ecological studies of invasive traits, applications of deep learning to plant identification, and recent advances in explainability methods;
- Chapter 3 presents the proposed methodology in detail, describing the classification model, the explainability pipeline, and the clustering process;
- Chapter 4 reports the results of experiments on the *Lythrum* dataset, including classification performance, analysis of salient regions, and discovery of morphological patterns;
- Chapter 5 concludes with a summary of contributions, a discussion of limitations, and suggestions for future directions, such as extending the pipeline to other taxa and integrating complementary trait data.

Finally, we identify two main contributions of this thesis:

- First, we demonstrated that identification of a species invasiveness from visual features alone is possible, without relying on tabular or categorical data;
- Second, we provided an interpretable framework to help future biological works to integrate ecological researches with computer vision, enhancing and reinforcing their results.

Chapter 2

Related Works

2.1 Biological approach to the identification of traits related to invasiveness

The study of traits that allow invasive plants species to succeed outside of their native range has received considerable attention in ecology, in order to advance towards successful prediction of invasiveness for different species of plants.

van Kleunen et al. (2015) [1] present a schema of questions to be asked regarding the success of alien species. The answer to one question is conditional on the answer to the previous ones (to account for the nested nature of the invasion process). The questions move from larger regions to smaller communities, and each one includes a series of traits that are likely to be related with the success of an alien species.

Other studies, such as Mathakutha et al. (2019) [2], explore the functional traits of invasive species, asking two major questions: (a) are invasive species functionally different or similar to native species? (b) which traits of invasives differ from traits of non-invasive aliens and thus confer invasibility? They state that most traits differed between invasive and native species, suggesting a correlation between functional traits and invasiveness. Additionally, they conclude that specific traits associated with invasiveness can be plant height, leaf area, frost tolerance and specific leaf area. A limitation of this work is the relatively small size of the sample for the study: they measured 13 traits for 26 species belonging to 13 different families that can be found in the sub-Antarctic region.

Another work that aims to identify determinants of plants invasiveness is the one by van Kleunen et al. (2010) [3]: they carry out a meta-analysis to study whether invasiveness is associated with performance-related traits (physiology, leaf-area allocation, shoot allocation, growth rate, size and fitness). They investigate pair-wise trait difference of 125 invasive and 196 non-invasive plant species, in the invasive range of the invasive species. They report that, for all traits, a greater

difference was shown between invasive and native species compared to invasive and other alien species. They also state that for comparisons between invasive species and native species that themselves are invasive elsewhere, no trait differences were significant. They conclude by suggesting that it might be possible to predict future plant invasions from species traits.

Li et al. (2024) [4] is another study that investigates the role of invasive-plant traits, native-plant traits, and their divergences in invasion processes. They suggest that the combination of invasive and native plants under study influences the results: for example, native plants such as *Artemisia argyi, Artemisia lavandulifolia* and *Chenopodium album* exhibited competitive superiority when co-occurring with the three invasive plants. *Setaria viridis, Austrocylindropuntia vestita*, and *Artemisia annua* had competitive superiority when they co-occurred with *Elodea canadensis, Galinsoga quadriradiata*, and *Erigeron annuus* respectively. Additionally they demonstrate that the competitiveness of invasive plants is mainly affected by height, diameter and biomass allocation, whereas native plants competitive abilities are primarily influenced by biomass allocation, diameter and function group differences.

Both [2] and [4] report about two hypothesis expressed by Ordonez et al. (2010) [5]: 'phenotypic divergence' and 'phenotypic convergence'. Phenotypic divergence proposed that successful invasive species possess traits different from native species, which allow them to better exploit empty niches. Phenotypic convergence instead comes from the idea that the environmental pressures limit the characteristics of species that can exist in an environment, resulting in similar traits between native and invasive species. Results from both [2] and [4] support the 'phenotypic divergence' hypothesis, suggesting that possessing different functional traits in alien species contributes to successful invasion.

A different point of view is proposed by Leffler et al. (2014) [6]: starting from the assumption that differences in trait values between native and alien invasive species may depend on the context of the comparison, they report and suggest that using trait values as predictors of future invasion will be a challenge. Instead they propose a criterion that differences in trait values between a native and exotic invasive species must be greater than differences between co-occurring natives for this difference to be ecologically meaningful. It is important to state that this work has been suggested to be flawed by Dawson et al. (2015) [7]: they state that the criterion proposed by Leffler et al. [6] cannot distinguish between cases where trait values may lie between those of native species but are still distinct and cases where they are very similar to native species.

These and other works provide useful insights to understand the differences between native species, non-invasive alien species and successfully invasive alien species, and to predict if a species has the ability to become invasive when introduced into a new region. However, despite proving that the invasion success of plants species can be predicted from functional traits, many of these studies suffer from

challenges and limitations, such as experimental conditions involving a limited number of plants or limited to a particular region, and are possibly difficult to propagate to bigger scales and with a greater amount of data. A question raises about the possibility of enabling this sort of analysis to scale using deep learning.

2.2 Deep learning for plant species identification

The creation of systems capable of recognizing plant species directly from images represents a significant interdisciplinary challenge that connects the fields of computer vision and biodiversity research. This task is part of fine-grained visual classification, where algorithms must differentiate among thousands of species by interpreting subtle morphological cues such as leaf form, venation patterns, or floral characteristics. The difficulty of this problem is amplified by the wide range of intraspecific variation caused by growth stage and environmental factors, as well as by the high degree of visual similarity shared among closely related taxa [8, 9, 10].

Initial methods for automated species recognition depended heavily on handcrafted features. For instance, systems like *Leafsnap* [11] employed classical computer vision strategies, including the extraction of curvature histograms from leaf outlines, to perform comparisons with reference databases. Although these approaches demonstrated that automated identification was feasible, they struggled to scale effectively and performed poorly on noisy, real-world images. The introduction of convolutional neural networks (CNNs) marked a turning point: by learning discriminative features directly from image pixels, CNNs rapidly surpassed traditional techniques and became the foundation of modern fine-grained plant classification. Because many botanical datasets are relatively small, researchers soon turned to transfer learning with pretrained models [12], which allowed for robust generalization even with limited training samples.

This advancement has been largely driven by the availability of large-scale, real-world datasets, among which the iNaturalist collection stands out. Introduced by Van Horn et al.[13], it includes over 859000 images spanning more than 5000 species contributed by a global network of citizen scientists, effectively capturing the visual diversity and ecological realism found in natural environments. The initial benchmark on this dataset achieved only 67% top-1 accuracy, with a significant drop in performance for rare species, revealing the inherent challenges posed by long-tailed distributions.

Herbarium specimens represent another valuable data source, offering a standardized yet taxonomically rich alternative for studying fine-grained classification. The Herbarium 2019 Challenge [14] showcased this potential by releasing a dataset of more than 46,000 labeled images from the Melastomataceae family across 683 species. In the associated FGVC6 competition, top-performing models achieved

89.8% classification accuracy. The subsequent Herbarium 2021 dataset [15] expanded dramatically to include over 2.5 million specimens representing 64,500 taxa. This dataset presented greater complexity due to a severe class imbalance (imbalance factor > 1650) and the inclusion of multiple major plant divisions. The corresponding competition evaluated models using the F1 score, and the top submission achieved 0.757. Carranza-Rojas et al. [16] further analyzed the application of CNNs to herbarium sheets, providing critical insights into the circumstances under which transfer learning can be beneficial or detrimental in this context.

The PlantCLEF series has been instrumental in advancing plant recognition research by posing increasingly demanding challenges. The 2022 edition [17] required identifying 80,000 species from 4 million images gathered from diverse data sources. Documentation highlighted the difficulty of developing models capable of generalizing across varying image qualities and acquisition types. In 2024, the competition [18] shifted toward identifying multiple species within vegetation plot images, reframing the problem as a weakly labeled multi-label classification task. Two pretrained baselines were released for this purpose, both based on the Vision Transformer (ViT) architecture originally trained with the DinoV2 self-supervised learning framework [19].

More recently, BioCLIP [20] introduced a large-scale foundation model trained on 10 million biological images using hierarchical contrastive learning with taxonomic supervision, setting a new benchmark in fine-grained biological classification. Its successor, BioCLIP 2 [21], further scaled this approach to 214 million images, significantly advancing the capabilities of foundation models in biological image understanding.

2.3 Deep learning for identification of traits related to invasiveness

Many studies approached the identification of invasive species through deep learning, to prevent and mitigate harm they might carry out to the environment.

Baron et al. (2018) [22] combines image processing and machine learning to identify yellow flag iris (*Iris pseudacorus*, an invasive species) from images obtained through a camera transported by a drone.

Similarly, Jensen et al. (2020) [23] employ a set of machine learning classifiers for detecting Kudzu vine (*Pueraria montana*, an invasive species) in the south-eastern area of the United States using spatial data.

Lake et al. (2022), instead, [24] used Worldview-2 and Planetscope satellite imagery to detect an invasive plant, leafy spurge (*Euphorbia virgata*), across complex landscapes using CNNs.

However, none of this research projects focus on species traits, but rather on

the identification of a known species in an unknown environment.

Moving towards the identification of traits belonging to invasive species, Keller et al. (2011) [25] evaluate trait-based risk assessments for invasive species using six diverse datasets (regional to global in scope), related to different taxa, regions and invasion stages. In these six datasets, two contain data for birds, two for fish, one for molluscs and one for pinus. For the latter (PinusG [26]), they took into consideration several categorical and numerical traits, identifying seed mass, dispersal mode, serotiny, generation time, reproductive intervals, fire tolerance, and environmental tolerances as key predictive traits. They compared two statistical methods and seven machine learning methods, without identifying any significant different result between the two approaches.

Evolving from this approach with the introduction of deep learning, in the latest years the concept of *imageomics* has emerged [27]. Works in this field aim to extract biological traits from images in different domains by introducing structured knowledge (from the biological domain) into deep learning algorithms: in particular, phenomics wants to extract phenotypical traits from image data [28, 29, 30].

In relation to this, Macleod (2017) [31] compares traditional geometric morphometric methods with newer machine-learning approaches for analyzing digital images of carnivore crania. It evaluates how well each method characterizes group differences and evaluates their suitability for morphometric analysis.

Lürig et al. (2018) [32] developed a pipeline to facilitate immediate extraction of high-dimensional phenotypic data from digital images, allowing biologists to focus on quick and reproducible collection of data.

Similarly, Porto et al. (2020) [33] propose a machine-learning-based pipeline to collect high-dimensional morphometric data in two-dimensional images of semi-rigid biological structures.

Previous research has mainly focused on numerical or tabular categorical traits (e.g. seed mass, dispersal mode), employing statistical or machine learning methods for the analysis. This approach may not capture morphological features that are exclusively visual, such as petal color or stem architecture.

At the same time, studies that applied deep learning mostly worked on the identification of known species instead of individual traits. To the best of our knowledge, no published study focuses on the identification of visual morphological traits that are related to the potential invasiveness of plants using deep learning.

2.4 Explainability mentions

The opaque or 'black box' nature of deep learning models remains a major obstacle to their broader use in ecological research, where both predictive accuracy and interpretability are essential. For these systems to be dependable, their outputs must

be explainable and grounded in biologically valid features rather than coincidental visual correlations. In fine-grained classification settings, this requires confirming that models focus on meaningful morphological traits, such as leaf outlines or floral structures, instead of unrelated image characteristics like illumination or background patterns. In important cases such as the detection of invasive species, interpretability ensures that predictions are based on diagnostic traits, reducing the probability of misleading conclusions.

Explainability approaches in computer vision can generally be grouped into three well-established categories:

- **Saliency and gradient-based methods.** Techniques such as CAM [34] and Grad-CAM [35] generate heatmaps that highlight image regions contributing most strongly to a prediction by computing class-specific gradients over the final convolutional layer. Integrated Gradients [36] takes a different approach by tracing a continuous path from a baseline image (for instance, a blank input) to the actual sample and integrating gradients along that trajectory to determine the relevance of each feature.

- **Perturbation-based methods.** These methods examine model sensitivity by directly modifying the input data. RISE [37] estimates the importance of different pixels by applying random masks and aggregating the model's responses. SHAP [38] calculates theoretically grounded importance scores for each input feature using Shapley values derived from cooperative game theory. LIME [39] interprets individual predictions by generating local perturbations around the input sample and fitting an interpretable surrogate model, such as a simple linear regression, to approximate the model's local decision boundary.

- **Concept-based and prototype-based methods.** These techniques aim to provide human-understandable reasoning that goes beyond feature attribution. TCAV [40] measures how much user-defined concepts influence model outputs. ProtoPNet [41], together with its improved variant Deformable ProtoPNet [42], learns prototypical parts during training, with final predictions determined by similarity to these learned prototypes. This design allows for explanations grounded in representative examples.

Such explainability frameworks have already demonstrated utility in ecological contexts. For example, in herbarium specimen classification, Grad-CAM visualizations revealed that networks often emulate the reasoning of taxonomic experts: they first assess the overall structure of the specimen and then attend to distinctive diagnostic regions, indicating that the model focuses on biologically relevant information [43]. More recent work employing concept-based methods in plant disease detection has uncovered both valid visual indicators and potential dataset biases [44]. Despite these encouraging developments, the consistent use

of explainable AI in biodiversity research (and particularly in invasive species detection and ecological monitoring) remains limited. Existing techniques still face several difficulties, including instability, a tendency to emphasize visually prominent yet biologically irrelevant patterns, and the absence of standardized quantitative metrics for validation. These limitations are even worse in ecological datasets, which often contain complex natural backgrounds and metadata leakage. In the present study, explainability is employed as a fundamental validation tool, ensuring that model decisions depend on genuine biological features and can therefore be considered trustworthy for scientific applications.

2.5 Research question

In conclusion, although several research project approached the identification of morphological traits related to plants invasiveness potential, we are not aware of any of them that relied exclusively on visual traits available from image data. By focusing on the plant genus *Lythrum* (Lythraceae) and by deploying deep learning methods with a systematic approach, we aim to differentiate invasive from non-invasive species through the recognition of their distinct morphological traits.

Chapter 3

Methods

3.1 Overview of the pipeline

In this section we present the pipeline for the methodology designed to extract meaningful information on morphological traits in invasive species.

The pipeline articulates in three main sections, each one containing intermediate steps:

1. **Classification model** to predict if an image is representing an invasive or non-invasive species
    - Extraction of **image embeddings** through a foundation model fine-tuned on the biological domain (BioCLIP 2);
    - **Classifier training** on the extracted embeddings;
2. **Explainability pipeline**
    - **Generation of heatmaps** to identify regions that drive the model predictions using Integrated Gradients as an XAI algorithm;
    - **Extraction of regions** from the heatmaps by selecting one or more bounding boxes that includes the most influential pixels;
    - **Clustering phase**
        - **Clustering of the regions** after embedding them individually and reducing them to a two-dimensional mapping with UMAP
        - **Cluster annotation** to determine the biological structure represented by each region (*Leaf, Flower, Stem*);
3. **Final analysis**

- **Pattern analysis** and discovery of visual traits that determine the potential invasiveness of a species;

In the following sections we will go into details for each step of the pipeline, explaining thoroughly the methodology and the decisions for the full research process.

3.2 Classification model

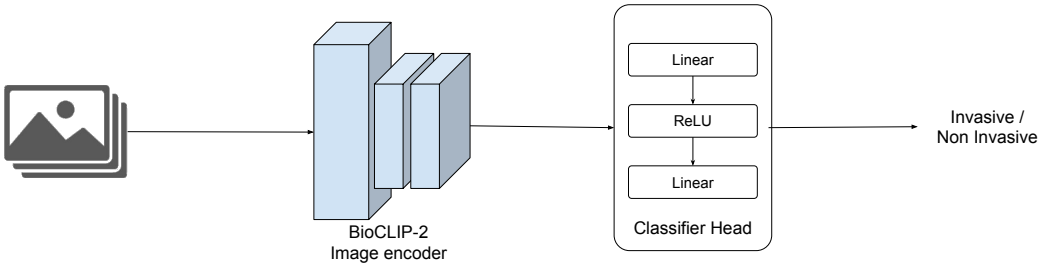


Figure 3.1: Classification model architecture.

In this section we describe how we built the model to take in input images from the dataset (4.1) and classify them as either *invasive* or *non-invasive*.

First we employed BioCLIP 2 [21] as an embedding extractor, by utilizing the image encoder of the model to extract a multidimensional embedding for every image in the dataset, which was then used to train a classifier. This image encoder was not finetuned on the dataset, meaning we only evaluated its capacity to extract a meaningful representation from the beginning. In this case any embedding extractor would have produced suitable embeddings, but BioCLIP is highly specialized in the biological domain and obtains better performances than a generic extractor, eliminating the need for further training or fine-tuning.

The classifier (which took as input the dimension of the embeddings and had a hidden dimension of 256) is comprised of a Linear layer, a ReLU activation layer and a final Linear layer with two possible outputs (invasive and non-invasive).

Finally, we trained the classifier to predict if the image (in the form of its embedding) is representing an *invasive* or *non-invasive* species, using cross-entropy as a loss function.

The complete schema with the architecture of the classification model is reported in Fig. 3.1.

To address the slight class imbalance in the dataset, class weights were applied in the cross-entropy loss. These weights were computed as the inverse of the logarithm

of the class sample counts, which provides a compromise between a better class balance and avoiding over-weighting less frequent categories.

The formula for cross-entropy loss and the computation of the weight for each class can be seen in Eq. 3.1.

$$\mathcal{L} = -\sum_{c=1}^{K} w_c y_c \log(\hat{y}_c), \qquad w_c = \frac{1}{\log\Big(1 + n_c\Big)} \tag{3.1}$$

The different models took into consideration as embedding extractors, the metrics used to compare them and the hyperparameters for the classifier are reported in Sec. 3.5.1.

The classification problem in our study is not trivial. Not all invasive species are invasive in every location or ecosystem: some studies do not consider native species as 'invasive', and choose to compare invasive alien species and native species, in order to study the traits that increase the chance of a non-native species to prevail over a native one. Other studies instead compare only non-native species, putting into contrast invasive aliens versus non-invasive aliens, addressing the question of what are the features that separate successful invaders and alien species that were not able to spread as successfully [1, 2, 3].

This study does not focus on correlation between species traits and geographical information, but only on the relation between a species appearence and its invasiveness potential. Therefore, we choose to look at this problem from a broader perspective, considering all species that are invasive somewhere as 'invasive' (at least potentially), and all species that are not invasive anywhere as 'non-invasive'. For this reason we ignore the 'native'/'non-native' comparison, focusing exclusively on the potential of a species to be invasive.

3.2.1 BioCLIP

In this section we describe what is BioCLIP and why we chose it (specifically BioCLIP 2) as a feature extractor model.

BioCLIP is a domain-specific vision-language foundation model developed to generalize across the entire tree of life [45]. This means that it is able to provide coverage across different taxa and to learn fine-grained representation of images of organisms which are often very similar to each other. It is also able to achieve strong performances in low data regimes, meaning it is able to produce useful embeddings even for species that are scarcely represented in the training data, or that are completely absent.

BioCLIP has been trained on TreeOfLife-10M [20], a large-scale diverse ML-ready biology image dataset, built integrating together different existing datasets

suchs as iNat21 [46], BIOSCAN-1M [47] and the Encyclopedia of Life[1].

Regarding the architecture, BioCLIP is built on OpenAI's CLIP framework [48], which relies on transformer architectures, using the *self-attention* mechanism to capture contextual relationships between elements in a sequence. In the vision domain, as in this case, this principle is applied through Vision Transformers (ViT), architectures that decompose an image into patches (rather than tokens, as would happen with a classic transformer), which are then serialized into vectors and processed by a transformer encoder as normal tokens. In BioCLIP, the vision encoder is a ViT-B/16, while the text encoder is a 77-token causal autoregressive transformer. Both map their inputs into a shared embedding space, in order to be able to measure similarity. The core of CLIP (and consequently BioCLIP) is the contrastive training objective: given a batch of paired image–text samples, the model learns to maximize the similarity between the embeddings of true pairs while minimizing it for mismatched pairs. This can be summarized as

$$\mathcal{L} = -\log \frac{\exp(\langle v_i, t_i \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle v_i, t_j \rangle / \tau)}, \tag{3.2}$$

where v_i and t_i are the normalized embeddings of the *i-th* image-text pair in the batch, and τ is a learned temperature parameter [48].

For BioCLIP, the text encoder receives as input different combinations and mixtures of common name, scientific name and taxonomic name (Tab. 3.1 reports an example of these combinations as presented in the original paper by Stevens et al. [20]). This multi-level textual supervision enriches the contrastive alignment process: not only does the model learn to associate an organism's image with a single label, but it also captures semantic structure across linguistic representations, increasing flexibility at testing time.

We chose to employ BioCLIP in our work because it is pre-trained (avoid the need for a long and computationally expensive training) on a vast dataset, which extensively covers the species we are taking into consideration. We also exploit its ability to learn fine-grained representation of images (in our case several species have very similar morphology) and for its ability to produce useful embeddings even in case of rarely represented species (as in our case, see Fig. 4.1).

In particular we chose the latest released version of BioCLIP, called BioCLIP 2 [21], which employs a more powerful vision transformer and is trained on a dataset with the same features of TreeOfLife-10M, but that extends to a much larger scale, called TreeOfLife-200M.

[1] https://eol.org

Text Type	Example
Common	black-billed magpie
Scientific	*Pica hudsonia*
Taxonomic	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia*
Scientific + Common	*Pica hudsonia* with common name black-billed magpie
Taxonomic + Common	*Animalia Chordata Aves Passeriformes Corvidae Pica hudsonia* with common name black-billed magpie

Table 3.1: Text types considered in the training of BioCLIP, as presented in the original paper [20].

3.3 Explainability pipeline

We implemented an explainability pipeline aimed at interpreting the decision-making process of our classification model. The objective was to verify whether the model's assessment of plant invasiveness relies on biologically relevant morphological traits, or whether it is unintentionally influenced by spurious correlations such as background elements. The procedure was organized into three main stages:

1. Generate heatmaps of model predictions to visualize the regions that most strongly influence classification outcomes;

2. Extract the image segments corresponding to those highlighted areas;

3. Assign biological labels to each extracted region to identify the structures being represented.

An overview of this process is presented in Fig. 3.2.

3.3.1 Heatmap generation

To obtain a complete understanding of the model's feature attributions, we employed two complementary heatmap generation techniques: **Integrated Gradients** [36] and **Gradient SHAP**, a stochastic variant of SHAP [38]. These approaches were selected to represent two distinct interpretability strategies: Integrated Gradients offers a deterministic, path-based gradient explanation, whereas Gradient SHAP incorporates randomness and connects attribution to Shapley value theory, ensuring equitable feature contribution estimates.

Integrated Gradients is a gradient-based saliency approach that quantifies pixel importance by integrating the gradients of the model's output with respect

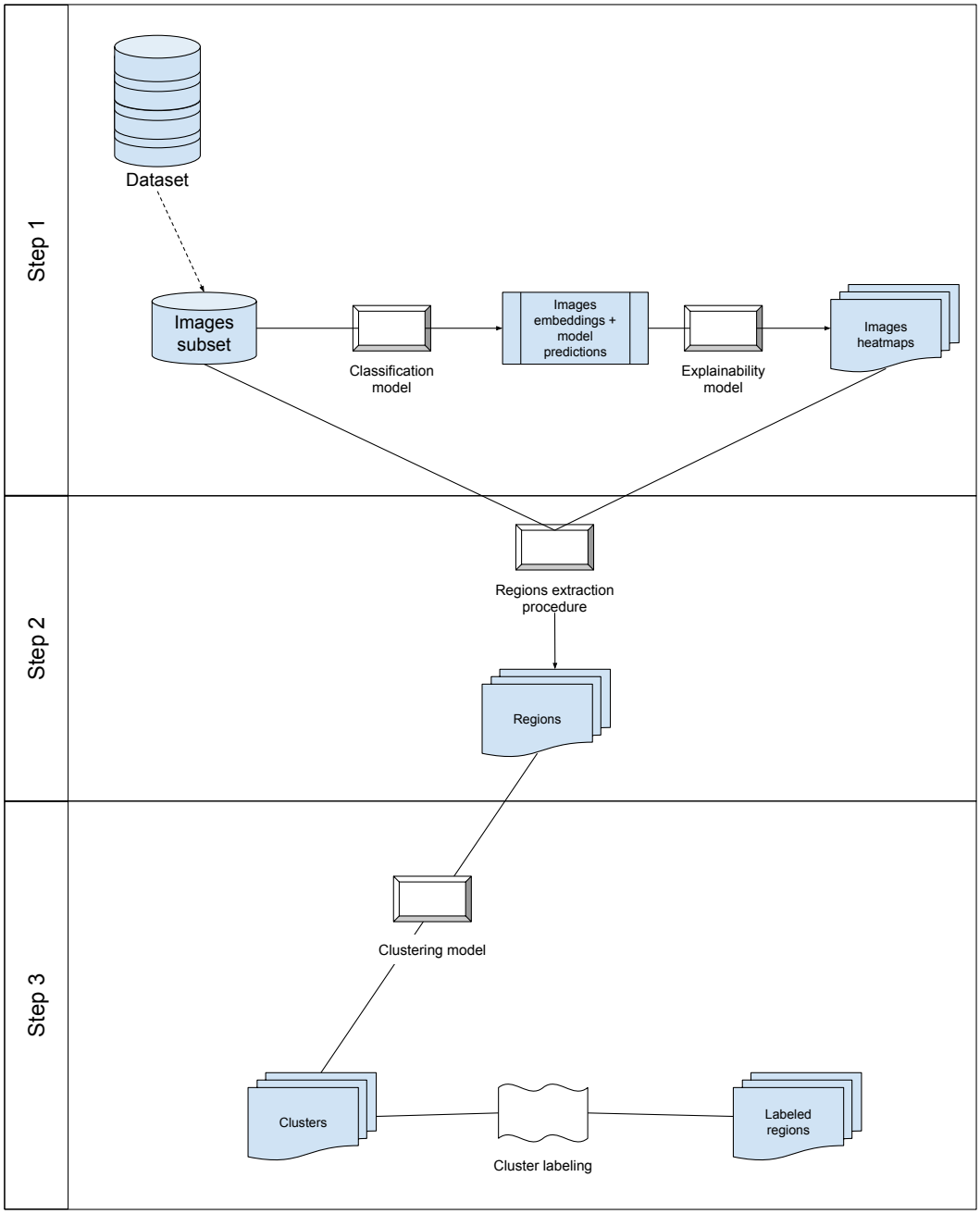


Figure 3.2: Explainability pipeline. The figure shows the three-step process: heatmap generation (top), extraction of relevant image regions (middle), and labeling of each extracted region (bottom).

to the input along a linear trajectory from a baseline image to the actual input. This method satisfies essential theoretical properties, including Sensitivity and Implementation Invariance, while requiring relatively few gradient computations. Conceptually, it measures each pixel's contribution to shifting the prediction away from the chosen baseline.

Gradient SHAP extends this concept by introducing noise to the baseline, creating multiple perturbed instances of it, and then averaging the integrated gradients across these samples. This procedure connects directly to the idea of Shapley values from cooperative game theory, ensuring a balanced attribution of influence across all input features.

Both methods were applied to the complete classification architecture, consisting of the frozen BioCLIP-2 image encoder followed by the classifier head. This stage corresponds to *Step 1* in Fig. 3.2. After comparing their performance based on clustering outcomes, one technique was selected as the final method for producing the definitive heatmaps.

3.3.2 Regions extraction

After producing the attribution maps, we implemented a systematic procedure to extract and store the most salient image regions (*Step 2* in Fig. 3.2). The complete workflow is illustrated in Fig. 3.3, which details each stage of the process:

1. **Normalization and preprocessing**: the original input tensor is reconstructed into an RGB image and normalized to the [0, 1] range. The attribution map is resized to the same spatial dimensions as the input image and normalized using the same scale. In cases where the attribution map contains multiple channels, it is converted to grayscale to simplify processing;

2. **Thresholding**: to retain only the most influential areas, we apply a percentile-based threshold. A binary mask is created by preserving pixels above the 90th percentile of attribution intensity, isolating the regions that contribute most strongly to the model's decision;

3. **Morphological processing**: to refine the binary mask and ensure clean, connected regions, we apply a sequence of morphological operations. The 'closing' operation fills small internal gaps and holes, while 'opening' eliminates isolated noise pixels that could otherwise interfere with later analysis;

4. **Connected components**: the refined mask is decomposed into connected regions, and for each component a bounding box is computed. Components with dimensions smaller than 10 px in either width or height are discarded as noise. Each valid bounding box is then expanded by 20 px in every direction to include the entire relevant morphological structure;

5. **Region cropping and saving**: finally, each selected region is cropped from the normalized RGB image and saved as an independent image segment. These cropped areas correspond to candidate biologically meaningful features, which will subsequently be clustered and annotated.

3.3.3 Clustering phase

After obtaining the heatmaps and extracting the salient image regions, the final step consisted of assigning a label to each region. The purpose of this stage was to identify whether a region represented a meaningful biological structure (and, if so, which specific one) or whether it instead captured irrelevant elements such as background patterns or human hands. This procedure corresponds to *Step 3* in Fig. 3.2.

Each extracted region was embedded into a semantic feature space using the image encoder of BioCLIP-2 [21], the same encoder integrated into our classification model (see Sec. 3.2). Because the resulting embeddings were high-dimensional, we applied **Uniform Manifold Approximation and Projection (UMAP)** [49] to reduce their dimensionality while maintaining local relationships among samples. UMAP constructs a high-dimensional graph that represents data topology and then optimizes a corresponding low-dimensional graph to preserve this structure as closely as possible, using cross-entropy as the similarity measure. This dimensionality reduction made the embeddings suitable for both visualization and clustering.

To group similar regions, we adopted a centroid-based clustering approach using the **KMeans** algorithm. This iterative method assigns each data point to the nearest cluster centroid, then recalculates centroid positions until convergence.

The outcome of this process was a mapping that linked each extracted image region to its respective cluster.

Finally, by visually inspecting the regions grouped within each cluster, we manually assigned descriptive biological or contextual labels: **Leaf**, **Flower**, **Stem**, **Hand**, and **Background/undefined**.

3.4 Final analysis

These initial steps set the foundation for the final phase of the analysis. To uncover meaningful patterns within the data, we developed the workflow depicted in Fig. 3.4, which is organized into three main stages:

1. **Region Extraction**: salient regions are extracted from all dataset images following the methodology described in Sec. 3.3.3;

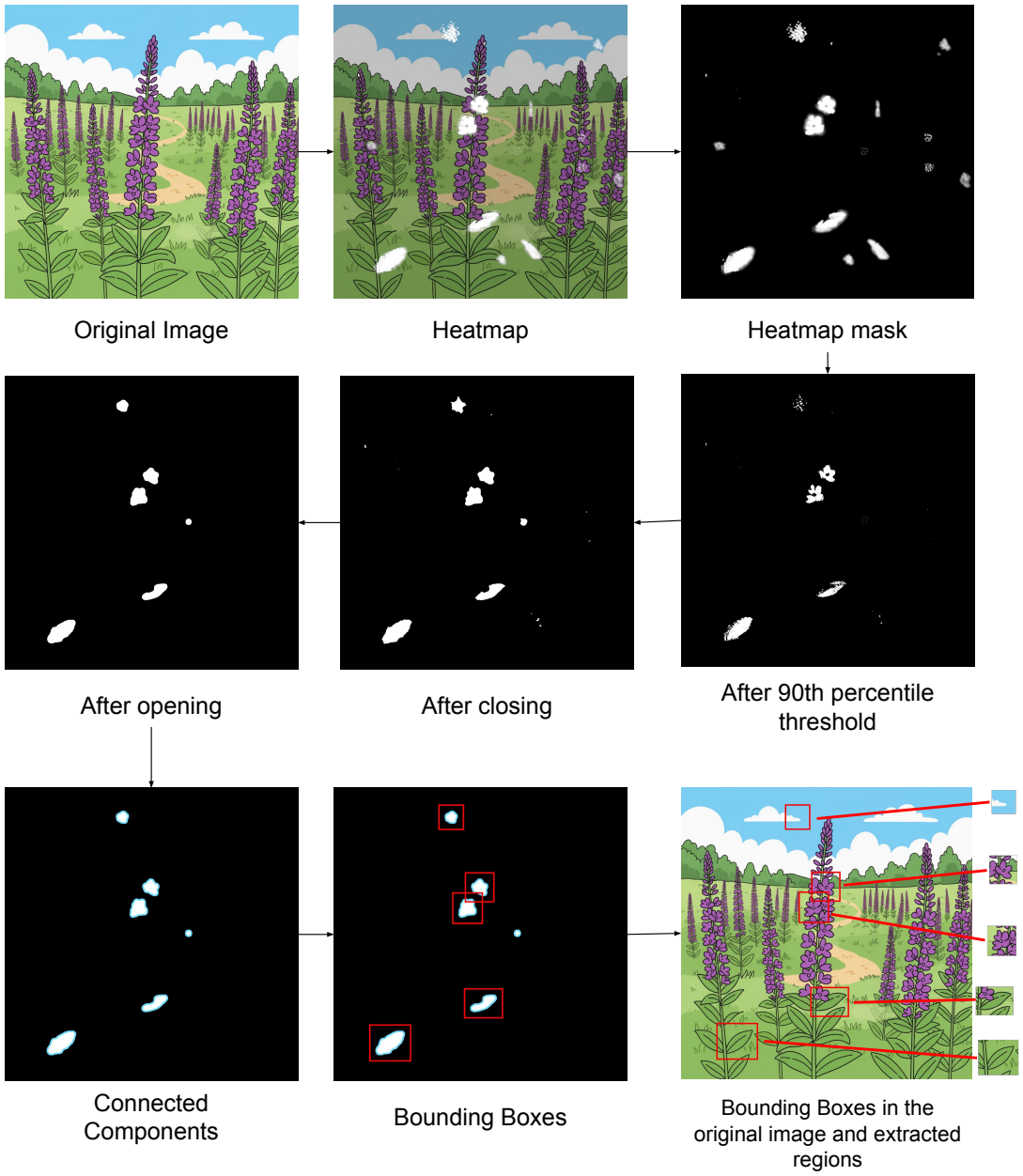


Figure 3.3: Details of the procedure that from an image and its heatmap extracts salient regions.

2. **Region Labeling**: the extracted regions are assigned to clusters using the KMeans model with fixed centroids obtained during training in Sec. 3.3.3. Since these clusters were manually annotated beforehand, each region automatically

inherits the label associated with its respective cluster;

3. **Pattern Analysis and Discovery**: as detailed in Sec. 3.4.1, we examine the distribution of labeled regions across the dataset to identify recurrent morphological structures and uncover patterns underlying model predictions.

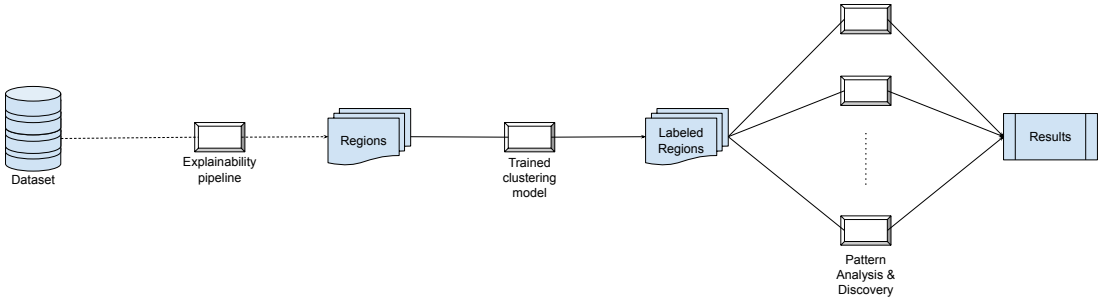


Figure 3.4: Final analysis pipeline. From the dataset, salient regions are extracted, clustered using the trained model described in Sec. 3.3.3, and labeled accordingly. After labeling, these regions become the ground for discovering and analyzing underlying prediction patterns.

For this last stage, we refined the cluster labels to achieve a more detailed biological interpretation. In addition to the general categories *Leaf*, *Flower*, and *Stem*, we incorporated characteristic traits that are specific to each species and structure, as described in Appendix A. The categories *Hand* and *Background/undefined* remained unchanged from their previous definitions.

3.4.1 Pattern Analysis and Discovery

Region-level information was aggregated to the image level to enable higher-level analysis. For each image, we computed the following variables:

- **Ground truth** and **predicted class**: both values were obtained directly from the classification model's outputs;
- **Prediction correctness**: a binary indicator specifying whether the predicted class matches the ground truth, along with the corresponding error category:
  - **True Positive (TP)**: invasive species correctly identified as invasive;
  - **True Negative (TN)**: non-invasive species correctly identified as non-invasive;
  - **False Positive (FP)**: non-invasive species incorrectly classified as invasive;

  - **False Negative (FN)**: invasive species incorrectly classified as non-invasive.

- **Region metrics**: these include the total number of extracted regions per image, the proportions of regions labeled as *Hand* and *Background/undefined*, and the coverage fraction, calculated as the ratio between the total area covered by the extracted regions and the full image area;

- **Trait metrics**: these capture the diversity and composition of biological traits within each image. Specifically, we measured the presence and relative frequency of each characteristic trait (as defined in Appendix A), the number of distinct traits (trait richness), and the Pielou evenness index, defined as

$$J' = \frac{H'}{\log S}$$

  where $H' = -\sum_{i=1}^{S} p_i \log p_i$ represents the Shannon diversity of the trait distribution, p_i is the relative frequency of trait i in the image, and S denotes the number of unique traits (richness). The value of J' ranges from 0, indicating a highly uneven distribution dominated by few traits, to 1, corresponding to traits perfectly evenly distributed in the image.

3.5 Experimental settings

3.5.1 Classification Model

For this work, three different models were evaluated as feature extractors for the images in the dataset:

- **ResNet18**: this Convolutional Neural Network (CNN) trained on ImageNet (a large dataset containing images of different objects) was chosen as a baseline, given its recognized quality in general-purpose image recognition [50].

- **BioCLIP 1**: BioCLIP is a contrastive learning vision model trained on TreeOfLife1-10M, a dataset of biological images of plants, animals and fungi [20].

- **BioCLIP 2**: BioCLIP 2 outperforms BioCLIP 1 by adopting a larger vision transformer and by training on a much larger and diverse dataset of images of plants, animals and fungi called TreeOfLife-200M [21].

The three models were compared as embedding extractors. Each model was used to extract a multidimensional embedding for every image in the dataset, which

was then used to train a classifier. For ResNet18 (imported from `torchvision`) the last Fully Connected layer was removed, whereas BioCLIP 1 and BioCLIP 2 where imported from HuggingFace and only the image encoder of the models was utilized. These embeddings extractors were not finetuned on the dataset, meaning we only evaluated their capacity to extract a meaningful representation without further training.

Once the embeddings were extracted and mapped to the original images, we split the dataset using 80% of the data to train the classifier, and 20% to test it on unseen data. The size and distribution of each split can be seen in Tab. 3.2

	Invasive	**Non-invasive**	**Total**
Training set	19898	16390	36288
Validation set	4935	4138	9073
Total	24733	20528	45361

Table 3.2: Support of the training set and validation set for invasive and non-invasive species

To evaluate the models, the classifier (Sec. 3.2) was trained for 50 epochs on the embeddings extracted by each model for the training data. At the end of every epoch it was evaluated on the unseen embeddings of the validation data. Early stopping (with no threshold for minimum improvement and a patience of 20 epochs) was implemented to prevent overfitting and to avoid wasting resources if the validation loss (the metric took into consideration for early stopping) was not improving.

The parameters used for the training of the classifier are reported in Tab. 3.3

Parameter	**Value**
Optimizer	Adam
Learning rate	1×10^{-4}
Loss function	Cross-entropy
Number of epochs	50
Batch size	32
Validation split	20%
Early stopping	Enabled
Random seed	42

Table 3.3: Training parameters used for the classification model. For the cross-entropy loss, we used as weights the reverse of the logarithms of the class samples, to contrast the slight imbalance in the class distribution.

To compare the different model we took into account the accuracy of the classifier

after the final epoch of training. We report the value for the different models in Tab. 3.4. As predictable, the best results are given by BioCLIP 2, which despite not being fine-tuned on our dataset and a arguably short training phase, already obtains satisfying results, proving to be a suitable model for our task (Fig. 3.5).

Model	Final accuracy	Final recall	Final F1 score	Final loss
ResNet18	0.779	0.78	0.78	0.483
BioCLIP 1	0.918	0.92	0.92	0.200
BioCLIP 2	**0.959**	**0.96**	**0.96**	**0.114**

Table 3.4: Results for the evaluation of the different models took into considerations, tested as feature extractors from the images. The values for the different metrics report the score for the evaluation after the last epoch of training.

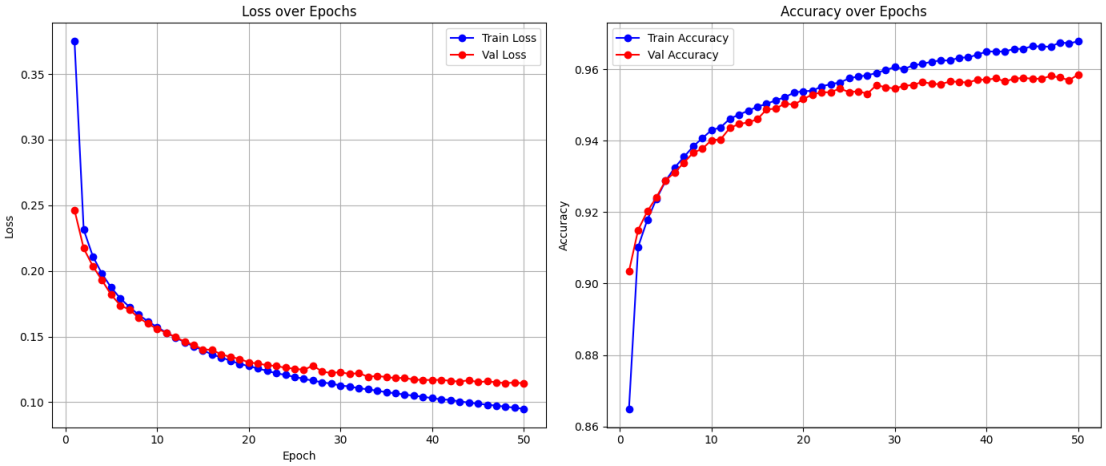


Figure 3.5: Values for the validation loss and validation accuracy throughout the training of the classifier using BioCLIP 2 embeddings. It can be seen how both metrics have not reached a plateau yet, suggesting that improvements on the results are possible with further training.

3.5.2 Explainability and Clustering phase

For both explainability methods, **Integrated Gradients** and **Gradient SHAP**, we relied on the implementations provided by the `Captum` library, while **UMAP** was applied using its dedicated Python package. In the case of Integrated Gradients, the path integral was approximated using 100 interpolation steps and an internal batch size of 5. For both explainability methods, the baseline was defined as a

completely black image, and the resulting pixel-wise attributions were aggregated across the RGB channels to produce a single normalized heatmap.

To determine the most suitable explainability method for our workflow, as well as the optimal hyperparameter configuration for both **KMeans** and **UMAP**, we conducted a systematic exploration of different parameter combinations. The full range of tested hyperparameters is summarized in Tab. 3.5. Each configuration was evaluated using the silhouette score computed on the resulting clustering assignments, as this metric provides a quantitative measure of both intra-cluster compactness and inter-cluster separation.

Hyperparameter		**Values**
Explainability method		Integrated Gradients, Gradient SHAP
KMeans	n_clusters	[30, 35, ..., 100]
	n_neighbors	[5, 10, 15, 20, 30, 50]
UMAP	min_dist	[0.01, 0.025, 0.05, 0.075, 0.1]
	distance_metric	euclidean, manhattan

Table 3.5: Hyperparameter configurations explored during the clustering phase.

Combining all the parameter values resulted in a total of 900 distinct configurations for each explainability method, leading to 1800 experiments in total. The clustering was carried out using the salient regions extracted from a subset of 2000 images: Integrated Gradients produced 7509 regions, whereas Gradient SHAP yielded 20021 regions.

Our results indicated that smaller values of the UMAP parameter *n_neighbors* (5–10) and lower *min_dist* settings (0.01–0.025) generally led to higher silhouette scores. Both explainability methods achieved comparable results, though Integrated Gradients consistently provided slightly better performance on average. The configuration with the best silhouette score was selected for subsequent analyses and is presented in Tab. 3.6, while its detailed results are reported in Tab. 3.7.

Explainability method	**n clusters**	**n neighbors**	**min dist**	**dist metric**
Integrated Gradients	30	5	0.025	manhattan

Table 3.6: Selected configuration for the clustering pipeline.

Cluster size				Silhouette
min	**max**	**mean**	**entropy**	**score**
88	441	250.3	0.984	0.428

Table 3.7: Clustering results for the configuration (Tab. 3.6) with the best silhouette score.

3.5.3 Final Analysis

Predictive Feature Analysis

We employed a **Random Forest** classifier to investigate which image-level features were most strongly associated with the correctness of model predictions. The input feature set included both the presence and frequency of traits, the total number of extracted regions, the count of distinct traits (richness), the coverage fraction, the Pielou evenness index, and the proportions of regions labeled as *Hand* and *Background/undefined*. Two complementary analyses were carried out:

1. a **global** analysis, aimed at ranking the features that distinguish correct from incorrect classifications;

2. an **error-type** analysis, performed in a one-vs-all scheme across the four prediction categories (TP, TN, FP, FN) to identify which features contribute to specific error patterns.

Feature importance was estimated using two distinct approaches. The first, *impurity-based importance*, evaluates the average reduction in node impurity attributed to each feature across all decision trees in the forest. This method provides a quick estimation of importance but tends to favor features with many possible values or continuous ranges, and may overemphasize features that create strong but localized splits. The second approach, *permutation-based importance*, measures the drop in model accuracy that occurs when the values of a single feature are randomly permuted. This metric captures each feature's actual predictive contribution and is less affected by biases related to scale or feature cardinality.

Metric-specific correlation with accuracy

To deepen the interpretation of the Random Forest results, we examined how specific image-level metrics correlate with prediction correctness. Each metric was discretized into bins, defined either by quantile ranges (for continuous variables) or by fixed-width intervals (for count-based metrics). For each bin, we computed the average prediction accuracy.

We also assessed the species composition within each bin by calculating the taxonomic distribution and comparing it to the overall dataset using the Kullback-Leibler (KL) divergence, defined as:

$$\mathrm{KL}(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)}$$

where $P(i)$ and $Q(i)$ represent the proportions of species i within the bin and in the overall dataset, respectively. This comparison enabled us to evaluate whether bins exhibiting distinct accuracy patterns shared a similar species composition. In cases where a bin's accuracy deviated substantially but its species distribution was highly unbalanced, the difference could be attributed to composition effects rather than the examined metric itself.

To statistically test for differences in accuracy among bins, we applied a one-way ANOVA. This method evaluates whether the mean prediction accuracy varies significantly across multiple groups (bins). In this context, it determines whether observed differences in correctness are genuinely linked to the binned metric, after accounting for within-bin variability.

Overall, this analysis provided a complementary perspective to the Random Forest feature importance results, allowing us to quantify the relationship between image-derived metrics and model performance while controlling for confounding effects of species composition.

Pairwise Trait Importance and Masked Image Analysis

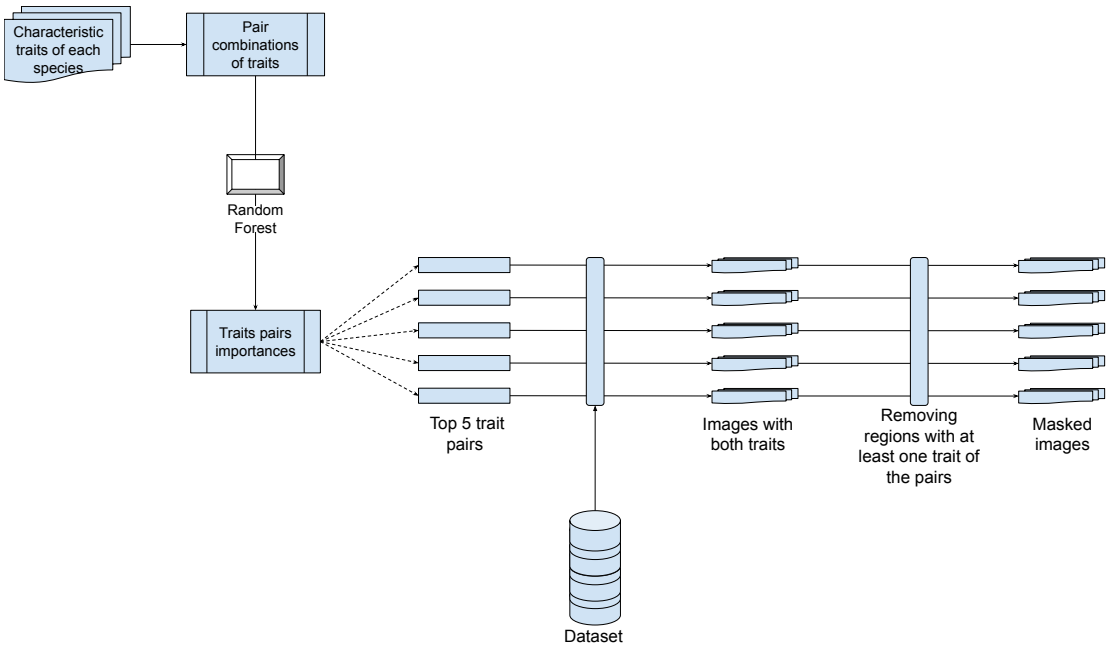


Figure 3.6: Creation of masked images: first the top 5 most important pairs of traits are identified. Then, images with both traits are selected (repeating the process for each pair of traits). For these images, we mask the regions containing at least one of the traits in the pair.

To explore how combinations of traits jointly influenced the model's predictions, we conducted a **pairwise predictive feature analysis** at the image level. The full procedure is illustrated in Fig. 3.6.

We first generated all possible pairs of characteristic traits, defined by species and plant structure (see Appendix A). These pairs were then categorized according to their taxonomic specificity into three groups: **Common** (traits shared by both invasive and non-invasive species), **Non-invasive only**, and **Invasive only**.

Using these pairwise features as predictors, we trained Random Forest classifiers to determine whether the model's original classification was correct. As in previous analyses, feature importance was computed both globally and separately for each prediction outcome (True Positive, True Negative, False Positive, False Negative). This approach made it possible to identify which trait combinations most strongly contributed to correct predictions or recurring errors. The five most important pairs were retained for further examination.

For each of the selected pairs, all images containing both traits were identified. The corresponding regions (those labeled with at least one of the two traits) were

then located within the original images and masked out by replacing them with transparent areas. These modified, or 'masked', images were reintroduced into the classification model to assess how accuracy and confidence changed when key trait information was removed.

By comparing model performance between pairs classified as *Common* and those marked as *Non-invasive only* (no *Invasive only* pairs were present in the dataset), we evaluated whether the classifier's decisions were driven primarily by general morphological cues shared among taxa or by features distinctive of non-invasive species.

Chapter 4

Results

4.1 Dataset construction

To build the dataset of species from the same genus, we first identified a particular species of interest. The selection fell on purple loosestrife (*Lythrum salicaria*), a species considered to be among the 100 world's worst invasive species [51], belonging to the family of Lythraceae.

We then retrieved all the species belonging to the *Lythrum* genus [52] (40 in total) and identified the ones that were invasive (3 in total: *Lythrum salicaria* [53, 54], *Lythrum hyssopifolia* [55, 56] and *Lythrum virgatum* [57, 58]).

To retrieve image data related to each species we used the `iNaturalist.org` website, an online social network where users can upload pictures of different living organisms from everywhere around the world.

iNaturalist provides different sets of APIs: first we made a sequence of calls to retrieve the iNaturalist ID for each manually retrieved taxon (e.g. *Lythrum virgatum*). We discarded the species for which we did not find a correspondence between taxon and id, meaning the species was not present on iNaturalist. Some species were present in the iNaturalist database, but had no available images: we discarded those from our dataset too. Then we used the taxon id to download every image present on iNaturalist (at download time, June 20th 2025) for the remaining 30 species [59]. The complete list of species used in our project and the correspondent number of images available for downloading can be seen in Fig.4.1.

Examples images taken from the dataset can be observed in Fig. 4.2.

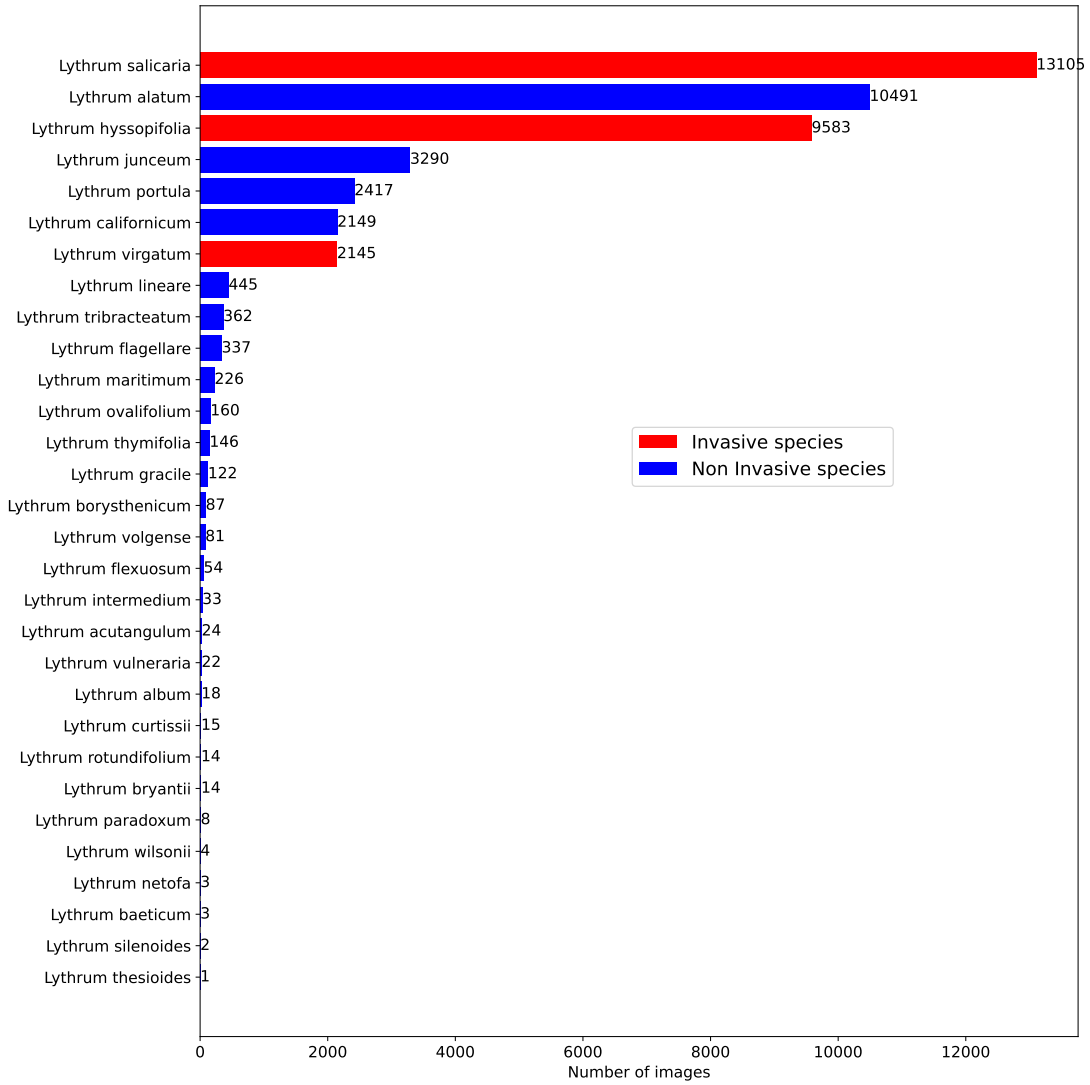


Figure 4.1: The distribution of the total number of images retrieved for each species. The invasive species are reported with a red column whereas the non-invasive species are reported in blue.

(a) *Lythrum salicaria* (Invasive)

(b) *Lythrum hyssopifolia* (Invasive)

(c) *Lythrum ovalifolium* (Non-invasive)

(d) *Lythrum album* (Non-invasive)

Figure 4.2: Examples of images representing different species inside the dataset, obtained from iNaturalist.

4.2 Classification model

4.2.1 Classification Model Cross Validation

To validate the accuracy and the quality of the prediction of the model (trained as described in Sec. 3.2) we decided to perform Leave One Species Out (LOSO) Cross Validation (Fig. 4.3). It consists in training 30 different models, one for every different species in the dataset, using as training data every image in the dataset except for the ones corresponding to a single species. The remaining elements of the dataset (all and only the images of that species) are used as a test set to evaluate the accuracy of the model.

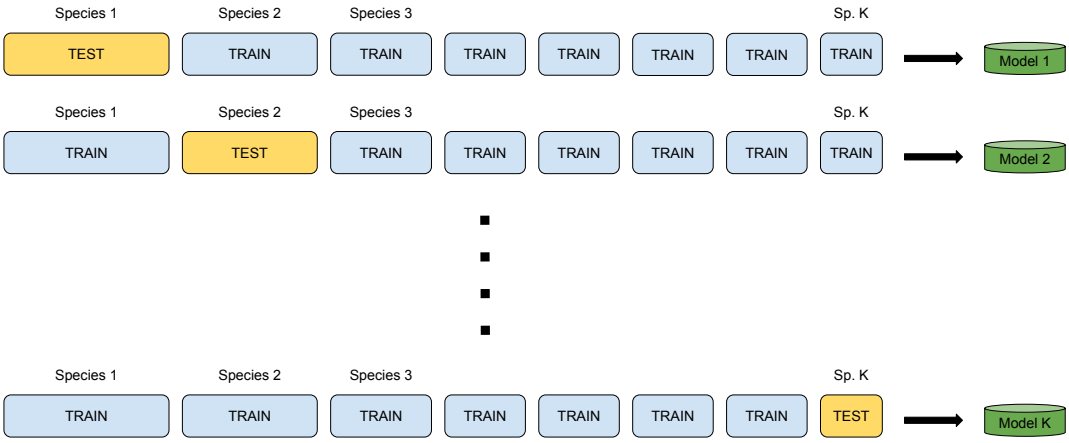


Figure 4.3: Leave One Species Out Cross Validation schema. Each iteration produces a model tested on a certain species, and is trained on all the dataset except for that species. Differently from K-Fold Cross Validation, each fold here represents a species, therefore they are not equivalent in samples size (see Tab. 4.1).

This allows us to explore the behavior of the model when encountering a species it has never seen before in the training phase, relying exclusively on the morphological traits of other species in the same genus for the predictions.

The results for the Leave One Species Out (LOSO) Cross Validation can be observed in Tab. 4.1.

There is a discrepancy between these results and the results for the accuracy of the classifier trained using all the species in a 80-20 split (Tab. 3.4). The average of the accuracies is lower than the overall accuracy of the model trained on all the species together, and the standard deviation shows instability in the results. Several species have mediocre accuracy (*L. salicaria, L. junceum, L. virgatum, L. tribracteatum*), few have very low accuracy (*L. thymifolia* and *L. wilsonii*, although

the sample size for *L. wilsonii* makes the result possibly not significant). For two species (*L. hyssopifolia, L. intermedium*) the model is completely or almost completely unable to classify it correctly. This suggests that the model, when seeing all the species at training time, might be learning to recognize the taxon and its classification instead of the traits and features that make a species invasive or not. In fact, when performing LOSO Cross Validation the model is forced to rely on morphological features for the prediction of invasiveness, and often obtains low or unsatisfying results with a taxon it has never seen in the training.

4.2.2 Mapping of the embeddings in a 2D space

To better study the difference in representation between the images of the different species, we applied Uniform Manifold Approximation and Projection (**UMAP**) to the multidimensional embeddings of the image data extracted with BioCLIP 2, reducing and mapping them to a 2-dimensional space (Fig. 4.4). The parameters used in this phase for UMAP can be found in Tab. 4.2

Additionally we calculated the average distance between each species (as the average distance between their points in the 2D space), to better understand which species were more visually similar and which species were more different. From this analysis we removed all species with less than 15 samples in the dataset, in order to have more meaningful comparisons. We report the average of all distances between species to be 6.584 ± 3.073.

First we analyze the results for *Lythrum salicaria* (the most represented species in our dataset, invasive) in Tab. 4.3: we can observe how *Lythrum intermedium*, a non-invasive species, is the closest to *L. salicaria*. Instead, the second closest is *Lythrum virgatum*, another invasive species, suggesting a visual similarity within species of the same class.

It is interesting to note how *Lythrum hyssopifolia*, one of the three invasive species and third most represented species in the dataset (Fig. 4.1), is the fourth-most distant species in the 2D representation, contrarily to what we would expect from our hypothesis. Its distance from other species (closest 5 and furthest 5) is reported in Tab. 4.4: the 5 closest species are all non-invasive despite *L. hyssopifolia* being invasive, whereas the other two invasive species are the second and third furthest species in this representation (the furthest being *Lythrum intermedium*).

Finally, we analyze *Lythrum intermedium* (Tab. 4.5), observing how *L. salicaria* and *L. virgatum* (both invasive) are the two closest species, whereas all the species in the bottom 5 for distance are non-invasive, with the exception for *L. hyssopifolia*. The average distance from *Lythrum intermedium* to other species is 11.332 ± 2.934.

It is important to indicate how some resources, such as the World Flora Online database [60] indicate *Lythrum intermedium* as a subspecies of *Lythrum salicaria*. This would explain the low distance between the two species, and therefore the

Species	Accuracy	Samples
Lythrum salicaria (I)	0.5138	13105
Lythrum alatum	0.8123	10491
Lythrum hyssopifolia (I)	0.0057	9583
Lythrum junceum	0.3173	3290
Lythrum portula	0.8250	2417
Lythrum californicum	0.8599	2149
Lythrum virgatum (I)	0.5888	2145
Lythrum lineare	0.9753	445
Lythrum tribracteatum	0.3343	362
Lythrum flagellare	0.8902	337
Lythrum maritimum	0.7743	226
Lythrum ovalifolium	0.7875	160
Lythrum thymifolia	0.1370	146
Lythrum gracile	0.8443	122
Lythrum borysthenicum	0.7816	87
Lythrum volgense	0.7037	81
Lythrum flexuosum	0.9630	54
Lythrum intermedium	0.0000	33
Lythrum acutangulum	0.8333	24
Lythrum vulneraria	1.0000	22
Lythrum album	1.0000	18
Lythrum curtissii	0.8667	15
Lythrum bryantii	1.0000	14
Lythrum rotundifolium	1.0000	14
Lythrum paradoxum	0.8750	8
Lythrum wilsonii	0.2500	4
Lythrum baeticum	1.0000	3
Lythrum netofa	1.0000	3
Lythrum silenoides	1.0000	2
Lythrum thesioides	1.0000	1
Mean ± Std	0.7313 ± 0.3088	–

Table 4.1: Model accuracy and sample sizes for Lythrum genus in the Leave One Species Out Cross Validation. Species indicated with (I) are invasive.

inability of the classifier to classify any sample of *L. intermedium* as non-invasive (see Sec. 4.2.1).

Parameter	Value
`n_neighbors`	15
`min_dist`	0.01
`metric`	euclidean
`random_state`	42

Table 4.2: Parameters used to map the embeddings in two dimensions with UMAP.

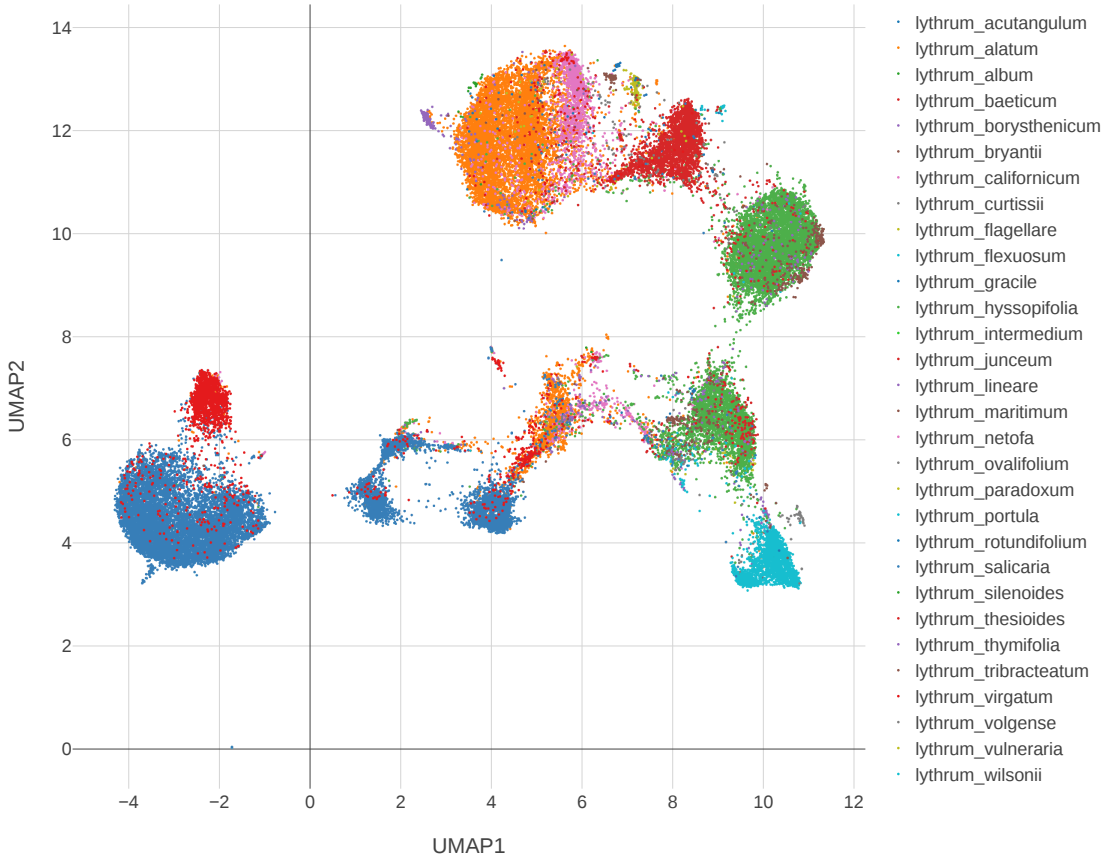


Figure 4.4: Two-dimensional UMAP projection of image embeddings. The invasive species are represented as follows: *Lythrum salicaria* (blue) occupies the region between UMAP1: −5 to 5 and UMAP2: 3 to 6.5. *Lythrum virgatum* (red) spans UMAP1: −5 to 6 and UMAP2: 4 to 8. *Lythrum hyssopifolia* (green) is located between UMAP1: 7–12 and UMAP2: 5 to 11.

Species 1	Species 2	Distance
lythrum salicaria (I)	lythrum intermedium	2.867
lythrum salicaria (I)	lythrum virgatum (I)	4.203
lythrum salicaria (I)	lythrum lineare	8.431
lythrum salicaria (I)	lythrum curtissii	8.885
lythrum salicaria (I)	lythrum alatum	8.891
lythrum salicaria (I)	lythrum junceum	11.512
lythrum salicaria (I)	lythrum hyssopifolia (I)	11.608
lythrum salicaria (I)	lythrum acutangulum	11.744
lythrum salicaria (I)	lythrum tribracteatum	11.801
lythrum salicaria (I)	lythrum flexuosum	12.386

Table 4.3: Top 5 closest and bottom 5 most distant Lythrum species pairs for *salicaria* (not including species with less than 15 total samples). Invasive species are indicated with (I).

Species 1	Species 2	Distance
Lythrum hyssopifolia (I)	Lythrum thymifolia	2.480
Lythrum hyssopifolia (I)	Lythrum tribracteatum	2.569
Lythrum hyssopifolia (I)	Lythrum junceum	3.921
Lythrum hyssopifolia (I)	Lythrum acutangulum	3.923
Lythrum hyssopifolia (I)	Lythrum flexuosum	4.045
Lythrum hyssopifolia (I)	Lythrum lineare	7.093
Lythrum hyssopifolia (I)	Lythrum album	7.197
Lythrum hyssopifolia (I)	Lythrum virgatum (I)	10.825
Lythrum hyssopifolia (I)	Lythrum salicaria (I)	11.608
Lythrum hyssopifolia (I)	Lythrum intermedium	13.321

Table 4.4: Top 5 closest and bottom 5 most distant species pairs for *hyssopifolia* (not including species with less than 15 total samples). Invasive species are indicated with (I).

4.2.3 Lythrum hyssopifolia exclusion

The results reported in Tab. 4.1, together with the distances reported in Tabs. 4.3 and 4.4 prompted us to reflect on the role of *Lythrum hyssopifolia* in the dataset.

L. hyssopifolia had an accuracy close to 0 in the Cross Validation experiments, possibly due to the far distance from other invasive species and the close distance to non-invasive species (Tab. 4.4). Similarly, *Lythrum salicaria* and *Lythrum virgatum* had mediocre accuracy, possibly due to the negative effect of them being far away from the third invasive species (which is also well represented in the dataset (Fig.

Species 1	Species 2	Distance
Lythrum intermedium	Lythrum salicaria (I)	2.867
Lythrum intermedium	Lythrum virgatum (I)	3.878
Lythrum intermedium	Lythrum lineare	9.413
Lythrum intermedium	Lythrum alatum	10.006
Lythrum intermedium	Lythrum curtissii	10.167
Lythrum intermedium	Lythrum volgense	13.155
Lythrum intermedium	Lythrum acutangulum	13.246
Lythrum intermedium	Lythrum hyssopifolia (I)	13.321
Lythrum intermedium	Lythrum tribracteatum	13.524
Lythrum intermedium	Lythrum flexuosum	13.870

Table 4.5: Top 5 closest and bottom 5 most distant species pairs for *intermedium* (not including species with less than 15 total samples). Invasive species are indicated with (I).

4.1). This effect is mitigated by the two species being close to each other (Tab. 4.3), a result which also seems to confirm our hypothesis that species with invasiveness potential share visual similarities.

Due to these results, we try to repeat the same LOSO Cross Validation but removing *Lythrum hyssopifolia* from the dataset, using only 29 folds in the validation experiment.

We can observe from Tab. 4.6 that removing *Lythrum hyssopifolia* from the dataset significantly improves the average accuracy between folds, reducing instability.

Looking into individual species, *Lythrum salicaria* increases its accuracy by 13.8%, whereas *L. virgatum* remains consistent with the previous accuracy. The results for these two species remain mediocre (although still better than the random choice threshold, for a two-classes problem), but this is partly explainable by the dataset slight imbalance: removing *L. hyssopifolia*, the dataset is comprised of 15150 samples labeled as 'invasive' and 20528 labeled as 'non-invasive'. For the LOSO Cross Validation, the fold that uses *L. salicaria* as a test set only includes samples for *Lythrum virgatum* for the 'invasive' class in the training set (2145 total), compared to 20528 samples for the 'non-invasive' class, making it very unbalanced. A similar analysis can be carried out for *Lythrum virgatum*, whose fold has 13105 samples labeled as 'invasive' (belonging to one species only), versus 20528 samples for the 'non-invasive' class.

This is a possible fault in the dataset, and must be taken into consideration when looking at the final results.

The biggest increase in accuracy can be seen in *L.junceum, L. tribracteatum*

Species	**Accuracy 1**	**Accuracy 2**	**Difference**	**Samples**
Lythrum salicaria (I)	0.5138	0.6520	+0.1382	13105
Lythrum alatum	0.8123	0.8485	+0.0362	10491
Lythrum hyssopifolia (I)	0.0057	-	-	9583
Lythrum junceum	0.3173	0.9675	+0.6502	3290
Lythrum portula	0.8250	0.9818	+0.1568	2489
Lythrum californicum	0.8599	0.9595	+0.0996	2149
Lythrum virgatum (I)	0.5888	0.5706	-0.0182	2145
Lythrum lineare	0.9753	0.9663	-0.0090	445
Lythrum tribracteatum	0.3343	0.9641	+0.6298	362
Lythrum flagellare	0.8902	1.0000	+0.1098	337
Lythrum maritimum	0.7743	0.9690	+0.1947	226
Lythrum ovalifolium	0.7875	0.9875	+0.2000	160
Lythrum thymifolia	0.1370	1.0000	+0.8630	159
Lythrum gracile	0.8443	0.9754	+0.1311	122
Lythrum borysthenicum	0.7816	0.9885	+0.2069	87
Lythrum volgense	0.7037	0.9630	+0.2593	81
Lythrum flexuosum	0.9630	0.9630	+0.0000	54
Lythrum intermedium	0.0000	0.0000	+0.0000	33
Lythrum acutangulum	0.8333	1.0000	+0.1667	30
Lythrum vulneraria	1.0000	1.0000	+0.0000	22
Lythrum album	1.0000	1.0000	+0.0000	18
Lythrum curtissii	0.8667	0.8667	+0.0000	15
Lythrum bryantii	1.0000	1.0000	+0.0000	14
Lythrum rotundifolium	1.0000	1.0000	+0.0000	14
Lythrum paradoxum	0.8750	1.0000	+0.1250	8
Lythrum wilsonii	0.2500	1.0000	+0.7500	4
Lythrum baeticum	1.0000	1.0000	+0.0000	3
Lythrum netofa	1.0000	1.0000	+0.0000	3
Lythrum silenoides	1.0000	1.0000	+0.0000	2
Lythrum thesioides	1.0000	1.0000	+0.0000	1
Mean ± Std	0.731 ± 0.31	0.929± 0.12	+0.198 ± 0.24	–

Table 4.6: Comparison of classification accuracy results for the LOSO Cross Validation of the model, when each fold includes *Lythrum hyssopifolia* in the training set (Accuracy 1) or not (Accuracy 2).

and *L. thymifolia* (*L. wilsonii* low sample size makes it difficult to carry out a significant analysis), the three species closest to *Lythrum hyssopifolia*: all of them are non-invasive.

This overall positive results validate the use of BioCLIP 2 as an embedding extractor, and our methodological approach for the training of the classifier. We then choose to identify *Lythrum hyssopifolia* as a possible anomaly or outlier in

the dataset, and we therefore remove this species from it.

Additionally, we choose to keep the 29 models (trained by excluding a single species every time from the training set) separated for the following analysis and experiments: this will allow us to obtain completely unbiased results, by working on every species with a model that has never seen it during the training phase.

4.3 Explainability pipeline

The explainability pipeline was described in Sec. 3.3 (see Fig. 3.2), whereas in this section we describe the results of the process.

4.3.1 Heatmap generation

To demonstrate the results produced by the explainability methods (*Step 1* in Fig. 3.2), Fig. 4.5 shows representative samples, each consisting of the original image, the generated heatmap, and the resulting overlay that highlights the most influential regions according to the model.

As seen in several examples, the attribution maps often align closely with biologically relevant structures, such as distinct plant organs or characteristic morphological features (Figs. 4.5a to 4.5d). In other instances, however, the highlighted areas do not correspond to meaningful biological traits. For example, in Fig. 4.5e, the method focuses on a human hand visible in the image, while in Fig. 4.5f, the heatmap is diffuse and fails to emphasize any clearly defined structure.

4.3.2 Regions extraction

For the first two examples displayed in Fig. 4.5, the results of the region extraction process (*Step 2* in Fig. 3.2) are shown in Figs. 4.6 and 4.7. A visual comparison reveals that Gradient SHAP (Figs. 4.6a and 4.7a) consistently produces a greater number of extracted patches than Integrated Gradients (Figs. 4.6b and 4.7b). This observation aligns with the quantitative results reported in Sec. 3.5.2, where Gradient SHAP generated 20021 regions from a sample of 2000 randomly selected images, while Integrated Gradients produced only 7509. For both techniques the extracted regions exhibit natural variation in size (and therefore in resolution) as expected from the extraction procedure.

A visual assessment alone does not provide a clear indication of which explainability method performs more effectively within the proposed pipeline. Therefore, the final choice was based on the results of the clustering hyperparameter optimization (see Sec. 3.5.2). As reported in Tab. 3.6, **Integrated Gradients** was ultimately selected as the preferred explainability approach.

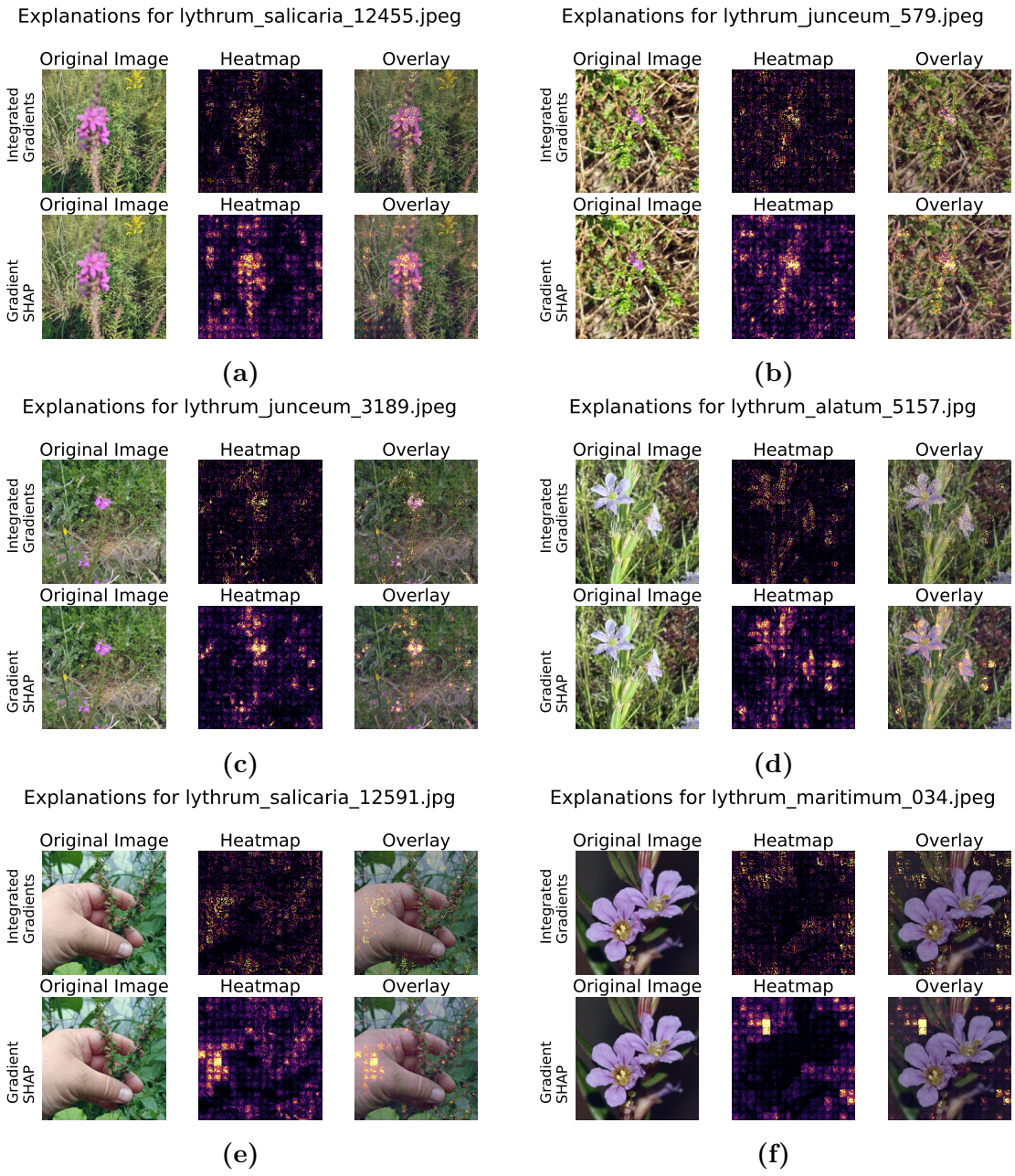
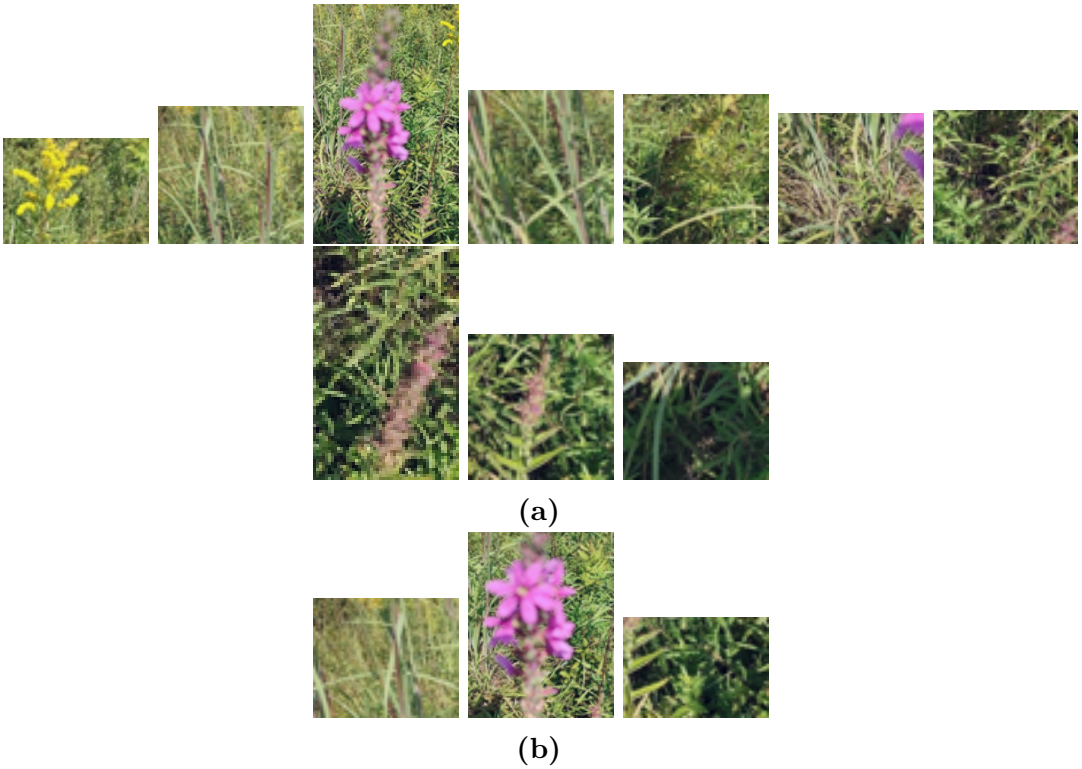


Figure 4.5: Examples of generated heatmaps. For each sample, the original image, the generated heatmap, and the overlay between the two are shown from left to right. Both Integrated Gradients (top row) and Gradient SHAP (bottom row) results are presented for each example. In most cases (a–d), the highlighted regions correspond to biologically meaningful structures, while in others (e–f) they do not align with the expected features.



(a)

(b)

Figure 4.6: Extracted regions for Fig. 4.5a by Gradient SHAP (a) and by Integrated Gradients (b).

4.3.3 Clustering phase

Sec. 3.3.3 described the implementation of the clustering phase (*Step 3* in Fig. 3.2); in this section, we analyze its results. The parameters used for this step are summarized in Tab. 3.6.

The clustering was first applied to a representative subset of 2000 images from the dataset. This preliminary stage allowed for manual inspection of the resulting clusters and for assigning descriptive labels based on visual examination. After labeling, the same cluster model (with fixed centroids) was applied to the full dataset during the final analysis (see Sec. 3.4). When using **Integrated Gradients** as the explainability method, a total of 7509 regions were extracted and divided into 30 clusters, as reported in Tab. 3.6.

We adopted **MiniBatchKMeans**, a scalable variant of KMeans that updates the centroids using small random batches of data rather than the entire dataset. This approach greatly reduces computational cost while maintaining similar clustering quality.

To interpret the clustering results, several complementary visualizations were

(a)

(b)

Figure 4.7: Extracted regions for Fig. 4.5b by Gradient SHAP (a) and by Integrated Gradients (b).

produced. The UMAP scatter plot in Fig. 4.8 shows the spatial distribution of embeddings together with their final cluster assignments. The bar chart in Fig. 4.9 presents the distribution of cluster sizes, highlighting the overall uniformity of the partition and supporting the cluster size entropy values reported in Tab. 3.7. Finally, Fig. 4.10 displays the average displacement of centroids across batches, showing that the movements gradually decrease and stabilize after a few iterations.

Cluster labeling

In this phase, clusters were manually annotated through visual inspection of the regions assigned to them. Each cluster received one or more labels describing the dominant visual content. For clarity, we organized the labels into two main categories:

- *Plant structures*:
  - **Leaf**: clusters primarly showing leaves;
  - **Flower**: clusters characterized by the presence of flowers;

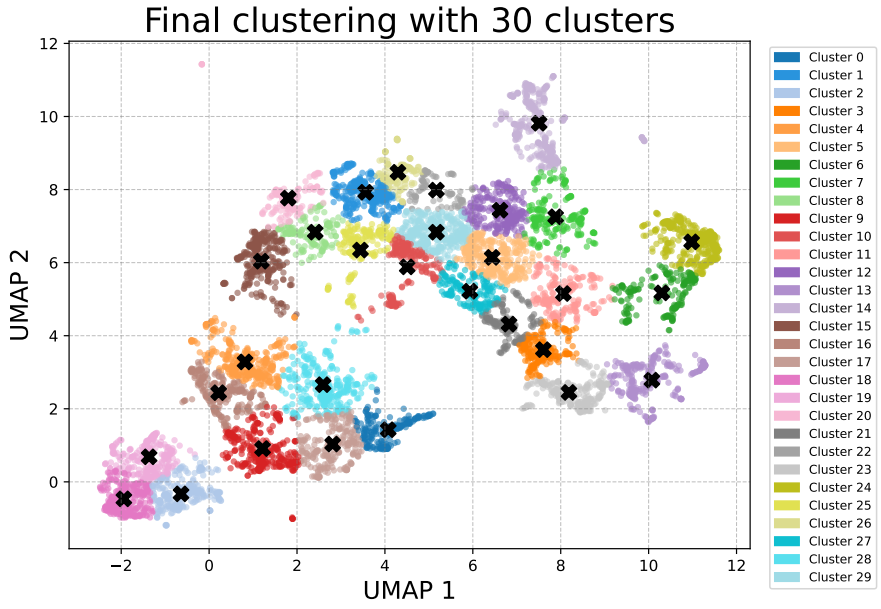


Figure 4.8: UMAP projection of the embeddings with final cluster assignments. Colors denote clusters and black markers indicate the final centroid positions.

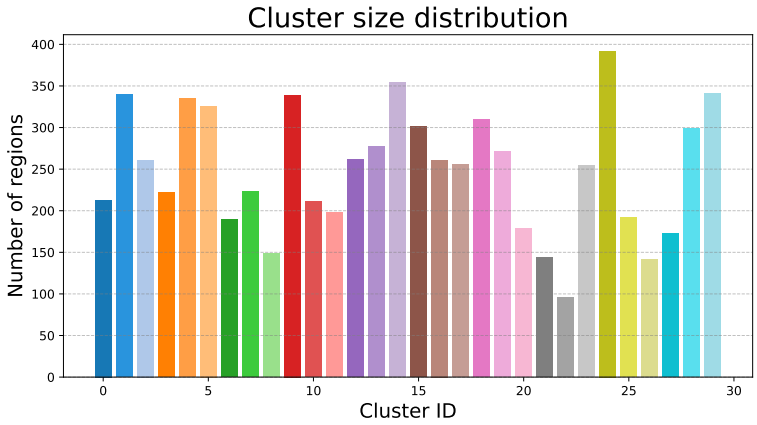


Figure 4.9: Distribution of cluster sizes for the final configuration. Colors match the corresponding clusters in Fig. 4.8.

  - **Stem**: clusters clearly depicting stems.

- *Spurious or non-informative features*:
  - **Hand**: clusters containing primarly parts of human hands or skin;
  - **Background/undefined**: clusters dominated by background fragments

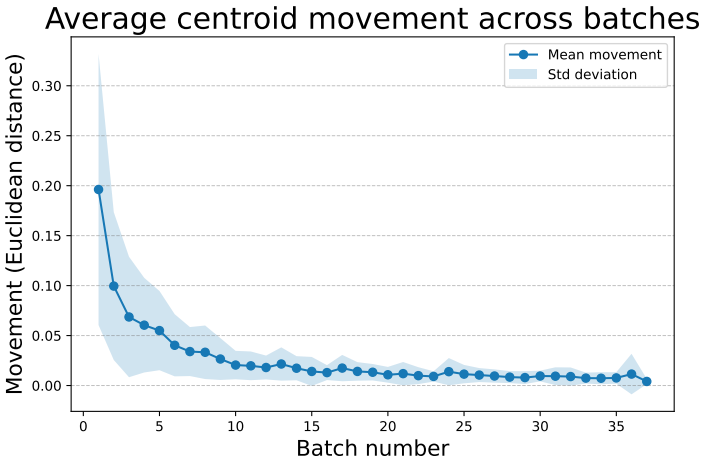


Figure 4.10: Average centroid displacement across minibatch updates. The line shows the mean movement across all centroids, the area represents the standard deviation.

or visually ambiguous regions.

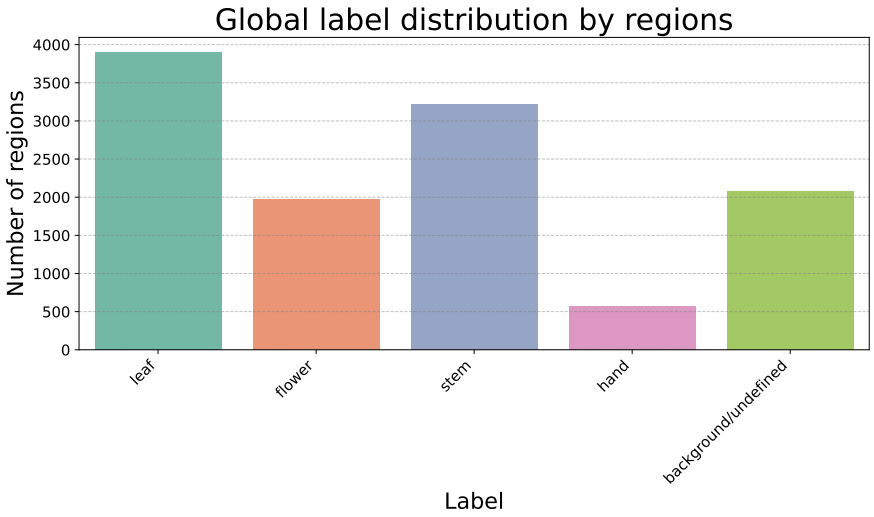
Labels within the same category are not mutually exclusive (e.g., a cluster may simultaneously be labeled *Leaf* and *Flower*), whereas labels from different categories are mutually exclusive. For instance, a cluster cannot be labeled both *Leaf* and *Hand*. This design reflects the aim of our analysis: to discriminate between clusters that captures biologically meaningful traits of the plants and clusters that corresponds to irrelevant or spurious visual cues.

Visualizations of labels distribution resulting from this phase are shown in Fig. 4.11. As expected, most extracted regions correspond to plant structure, with *Leaf* being the most frequent. Non-biological features are still a relevant number, expecially those labeled *Background/undefined*.

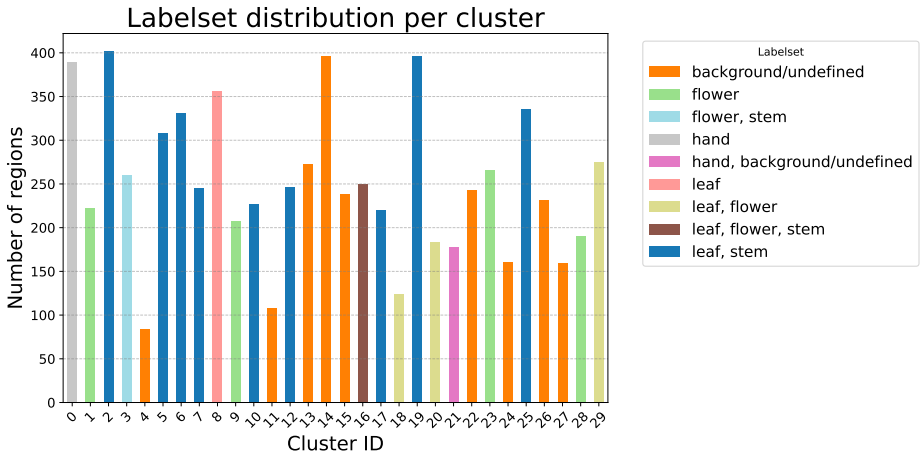
Examples of regions assigned to different clusters, with the correspondent labeling, can be observed in Fig.4.12, 4.13, 4.14, 4.15.

Clustering validation

To assess the reliability of both the KMeans clustering results and the corresponding manual annotations, we performed an additional validation using the density-based clustering algorithm **HDBSCAN**. Using the same set of extracted regions, we re-clustered the data and examined how HDBSCAN grouped the samples, focusing on the consistency of its partitions with the manually assigned labels. The procedure followed the same workflow previously used for KMeans (see Sec. 4.3.3): extracted regions were first embedded using the BioCLIP-2 image encoder, then



(a)



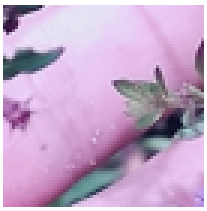
(b)

Figure 4.11: Results of manual cluster labeling: global distribution by number of regions (a) and labelset assigned to each cluster (b) are shown.

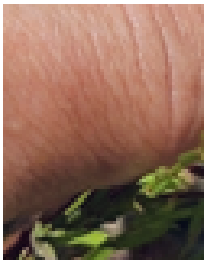
reduced in dimensionality with UMAP, and finally clustered. For consistency, the regions produced with Integrated Gradients were employed, and the same UMAP parameters reported in Tab. 3.6 were used. In HDBSCAN, the main hyperparameter is the minimum cluster size, which we varied across the range

$$min\_cluster\_size = [5, 6, \ldots, 71].$$

To ensure a fair comparison with KMeans, we excluded configurations that produced fewer than 10 clusters or more than 100. As a result, out of the 66 tested



(a) *Lythrum alatum*

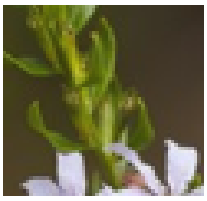


(b) *Lythrum californicum*

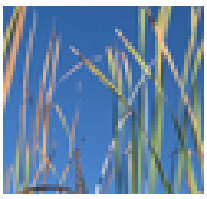
(c) *Lythrum alatum*

Figure 4.12: Three of the regions present in cluster with id=0, which was assigned the label **Hand**. Each region reports the species represented in the original image.

(a) *Lythrum junceum*



(b) *Lythrum alatum*

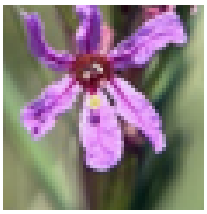


(c) *Lythrum salicaria*

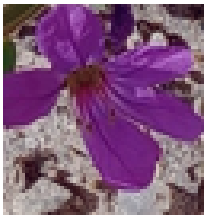
Figure 4.13: Three of the regions present in cluster with id=16, which was assigned the label **Leaf, Flower, Stem**. Each region reports the species represented in the original image.

configurations, only 9 satisfied this condition. This finding is expected given that HDBSCAN operates under a fundamentally different clustering principle: whereas KMeans partitions the data into a fixed number of clusters, HDBSCAN identifies clusters of varying density and assigns low-density points to a noise category. Despite the smaller number of valid configurations, this subset was sufficient for our purpose. The objective here was not to fine-tune HDBSCAN but to evaluate the robustness of the KMeans-based clustering and the coherence of our manual labeling. Even a limited number of well-defined, data-driven configurations can provide a solid validation basis.

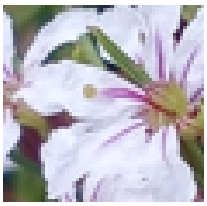
Fig. 4.16 summarizes the validation results. For each valid HDBSCAN configuration, we calculated the *cluster label consistency ratio*, defined as the proportion of samples in each cluster that belong to its dominant label set. This measure quantifies cluster purity with respect to the manually assigned labels. Across all retained configurations, the mean consistency ratio was 0.75, with the best-performing configuration ($min\_cluster\_size = 16$) achieving a value of 0.92 (see Fig. 4.16b). As shown in Fig. 4.16a, most clusters in the tested configurations reached a consistency ratio of 1.0, indicating strong alignment with the labeling



(a) *Lythrum californicum*



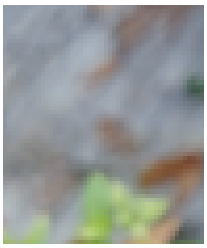
(b) *Lythrum junceum*



(c) *Lythrum lineare*

Figure 4.14: Three of the regions present in cluster with id=23, which was assigned the label **Flower**. Each region reports the species represented in the original image.

(a) *Lythrum californicum*



(b) *Lythrum salicaria*

(c) *Lythrum portula*

Figure 4.15: Three of the regions present in cluster with id=27, which was assigned the label **Background/undefined**. Each region reports the species represented in the original image.

produced by our pipeline. A more detailed discussion of this validation analysis is presented in Appendix B.

Overall, these results demonstrate that our manually assigned labels reflect the internal structure of the data and are consistent with a density-based clustering approach.

4.4 Final analysis

Fig. 3.4 presents the workflow of the final analysis, as introduced in Sec. 3.4. In this section, we report and discuss the corresponding results.

The complete dataset includes 35678 images. Using the proposed extraction pipeline (Sec. 3.3.2), we obtained a total of 132128 regions. The results of the following clustering stage are summarized in Fig. 4.17. As illustrated in Fig. 4.17a, each extracted region inherits the label set of the cluster to which it has been assigned. The overall distributions of labels and label sets for all extracted regions are shown in Figs. 4.17b and 4.17c, respectively.

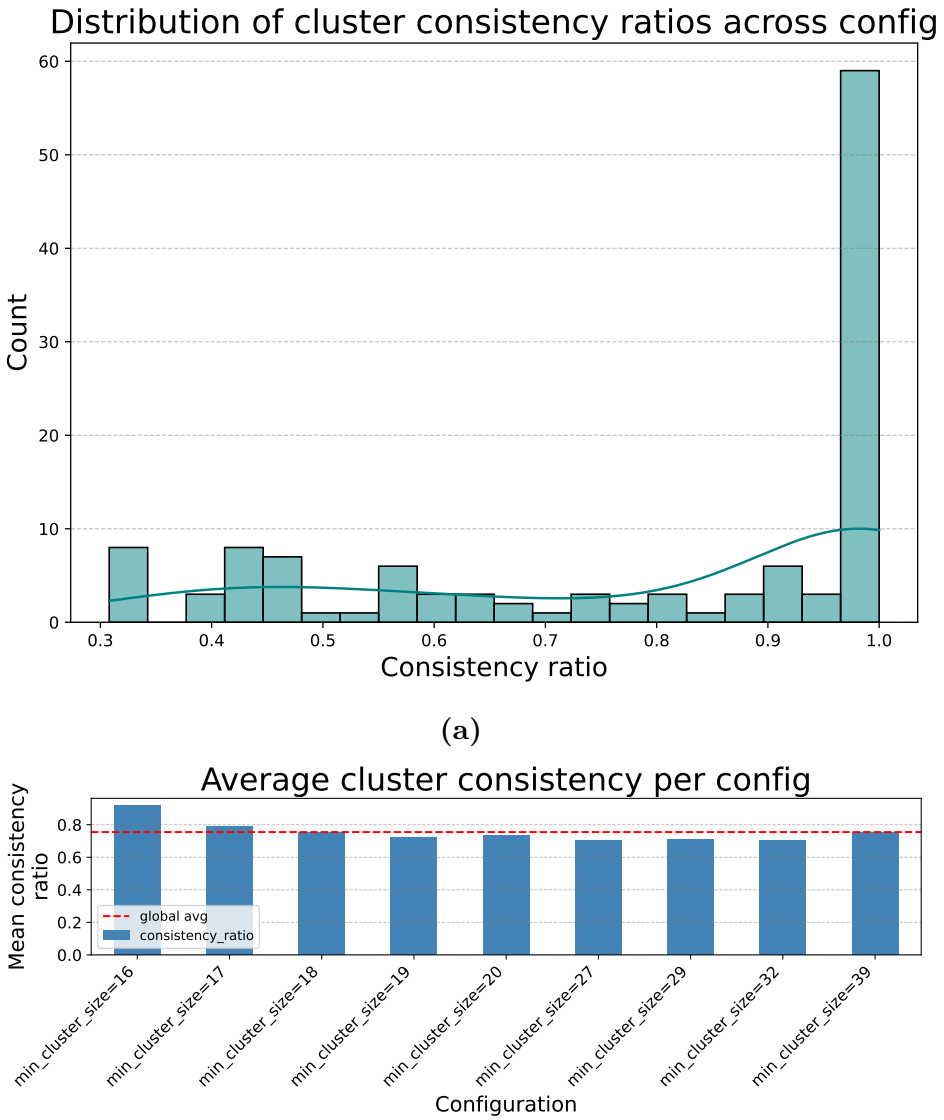


Figure 4.16: Clustering and labeling validation using HDBSCAN. For considered configurations, the distribution of cluster consistency ratio (a) and the average cluster consistency per configuration (b) is shown.

4.4.1 Predictive Feature Analysis

As outlined in Sec. 3.5.3, we conducted a predictive feature analysis using Random Forests to evaluate the importance of various features, including both the characteristic traits described in Appendix A and the additional metrics computed

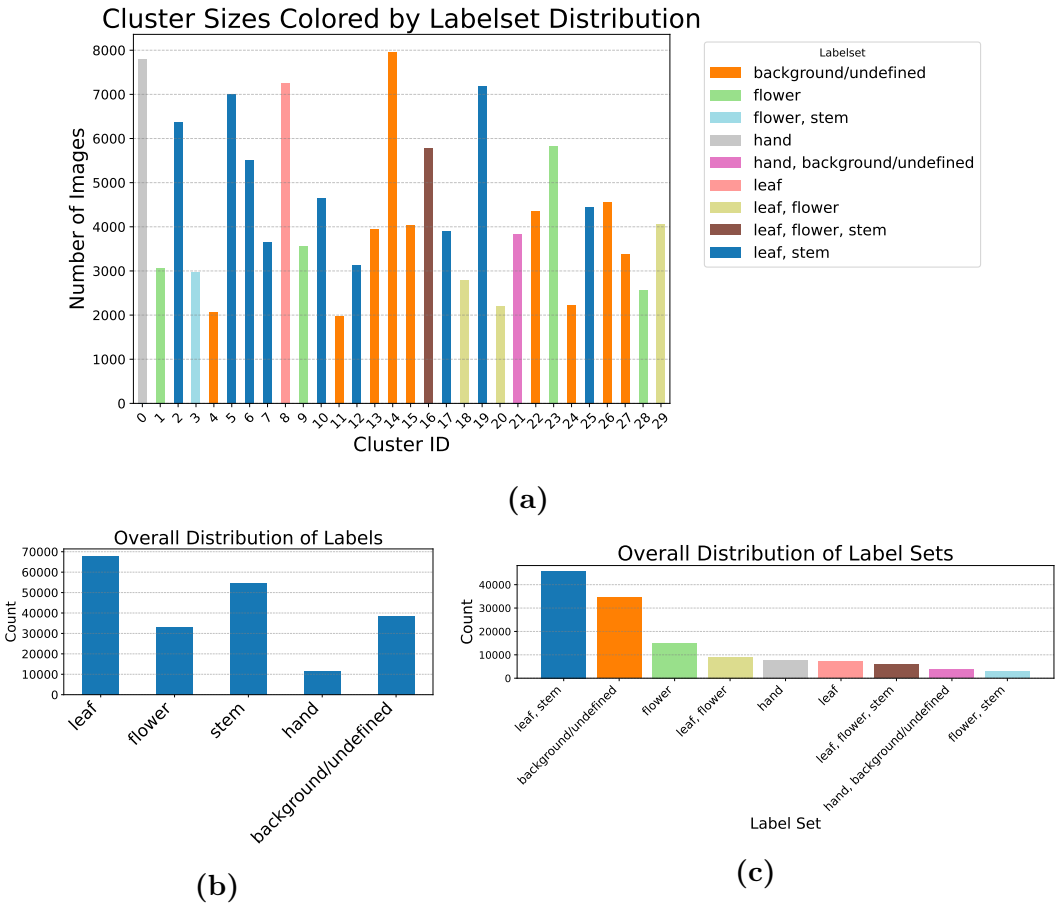


Figure 4.17: Results of the clustering and labeling of regions extracted from the entire dataset. In particular, cluster sizes distribution colored by the corresponding labelset of each cluster (a), labels distribution across all clusters (b) and labelset distribution across all clusters (c) are shown.

as detailed in Sec. 3.4.1. The results of this analysis are presented in Tabs. 4.7 to 4.9, corresponding respectively to the global analysis, the True Positive and True Negative analyses, and the False Positive and False Negative analyses.

Tab. 4.7 presents the results of the global feature importance analysis for correct predictions across the entire dataset. The fraction of the image covered by extracted regions (`coverage_frac`) emerges as the most influential predictor (impurity importance = 0.6198; permutation importance = 0.2224), suggesting that images with higher region coverage are considerably more likely to be classified correctly. Secondary predictors include `pielou_evenness`, `background_frac`, and `hand_frac`, which contribute moderately to the overall accuracy. While permutation importance values are generally lower than impurity-based scores, they confirm the same

Feature	Impurity	Permutation	
	Importance	Importance	Std
coverage_frac	0.6198	0.2224	0.0017
pielou_evenness	0.4101	0.0557	0.0007
background_frac	0.0329	0.0729	0.0009
hand_frac	0.0314	0.0571	0.0009
trait_erect_freq	0.0287	0.0545	0.0005
trait_rounded_at_the_base_freq	0.0216	0.0387	0.0009
richness	0.0021	0.0486	0.0008
trait_rounded_at_the_base_present	0.0152	-	-
n_regions	0.0149	0.0564	0.0008
trait_sessile_freq	0.0125	0.0304	0.0007

Table 4.7: Global feature importance analysis from Random Forest classification. The table reports both impurity based importance and permutation importance (accompanied by the standard deviation across 20 repetitions) for each feature. A dash (-) indicates that the value was not computed or not meaningful for that feature, typically due to low prevalence or insufficient variation in the subset.

ranking among the top predictors. Notably, `richness` and `n_regions` show relatively low impurity importance but moderate permutation importance, implying that these features may interact with others to influence model performance.

In summary, the global analysis highlights the most impactful individual predictors, though it may conceal more nuanced feature-specific effects that differ across prediction categories.

Error-Type Analysis

To better capture specific relationships between features and prediction outcomes, we conducted separate analyses for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) cases (Tabs. 4.8 and 4.9). This approach allowed us to identify which features contribute to different types of correct and incorrect predictions, providing a more detailed understanding than the global analysis alone.

For TP and TN outcomes, `coverage_frac` remains the most influential predictor (TP: 0.4099 impurity, 0.1642 permutation; TN: 0.1650 impurity, 0.0881 permutation), confirming its strong role in determining correct classifications. Traits associated with leaf base morphology (`rounded_at_the_base`) and flower morphology (`flowers_in_whorled_clusters`) also display higher importance, suggesting that these features contribute to differentiating between true positive and true negative predictions. The general consistency between impurity-based and

Feature	Impurity Importance		Permutation Importance		Permutation Std	
	TP	**TN**	**TP**	**TN**	**TP**	**TN**
coverage_frac	0.4099	0.1650	0.1642	0.0881	0.0013	0.0011
trait_rounded_at the_base_freq	0.0652	0.1015	0.0300	-	0.0006	-
trait_rounded_at the_base_present	0.0613	0.0981	0.0219	-	0.0006	-
trait_flowers_in whorled_clusters_freq	0.0331	0.0542	-	-	-	-
trait_flowers_in_ _whorled_clusters_present	0.0278	0.0426	-	-	-	-
pielou_evenness	0.0272	0.0327	0.0294	0.0327	0.0006	0.0004
trait_attenuate_at_ _the_base_freq	0.0219	0.0320	-	0.0205	-	0.0003
background_frac	0.0218	-	0.0508	0.0446	0.0006	0.0006
trait_erect_freq	0.0209	-	0.0325	0.0220	0.0005	0.0005
hand_frac	0.0193	-	0.0433	0.0276	0.0007	0.0005
richness	0.0185	0.0186	0.0273	0.0345	0.0005	0.0005
n_regions	-	-	0.0413	0.0352	0.0006	0.0008
trait_sessile_freq	-	-	0.0214	0.0222	0.0006	0.0006
trait_opposite_freq	-	-	-	0.0148	-	0.0003
trait_floral_tube cylindrical_freq	-	-	-	0.0147	-	0.0004

Table 4.8: Feature importance analysis from Random Forest for True Positive (TP) and True Negative (TN) classifications. The table reports both impurity based importance and permutation importance (accompanied by the standard deviation across 20 repetitions) for each feature. A dash (-) indicates that the value was not computed or not meaningful for that feature, typically due to low prevalence or insufficient variation in the subset.

permutation-based rankings supports the robustness of these findings. Although both TP and TN analyses reveal similar patterns, TP cases depend more heavily on `coverage_frac`, while TN cases show slightly greater influence from specific morphological traits. This indicates subtle differences in the features that drive accurate positive versus accurate negative predictions.

For FP and FN outcomes, `coverage_frac` again emerges as the dominant factor (FP: 0.6493 impurity, 0.0684 permutation; FN: 0.5372 impurity, 0.1558 permutation). Additional contributors include `pielou_evenness`, `background_frac`,

Feature	Impurity Importance		Permutation Importance		Permutation Std	
	FP	**FN**	**FP**	**FN**	**FP**	**FN**
coverage_frac	0.6493	0.5372	0.0684	0.1558	0.0008	0.0014
pielou_evenness	0.0470	0.0344	0.0287	0.0301	0.0004	0.0007
background_frac	0.0357	0.0279	0.0263	0.0482	0.0004	0.0005
hand_frac	0.0344	0.0266	0.0177	0.0400	0.0005	0.0006
trait_erect_freq	0.0274	0.0262	0.0217	0.0313	0.0004	0.0005
trait_attenuate_at the_base_freq	0.0253	0.0128	0.0305	0.0074	0.0005	0.0001
richness	0.0218	0.0198	0.0249	0.0294	0.0005	0.0005
n_regions	0.0172	0.0124	0.0195	0.0393	0.0004	0.0008
trait_rounded_at the_base_freq	-	0.0403	-	0.0333	-	0.0007
trait_rounded_at the_base_present	-	0.0341	-	0.0202	-	0.0005
trait_sessile_freq	0.0102	0.0120	0.0176	0.0166	0.0004	0.0007
trait_opposite_freq	0.0106	-	0.0171	-	0.0004	-
trait_opposite, becoming_alternate distally_freq	0.0116	-	0.0215	0.0122	0.0005	0.0006
trait_stamens_6_freq	0.0125	-	0.0198	-	0.0003	-
trait_floral_tube cylindrical_freq	0.0087	-	0.0134	-	0.0003	-

Table 4.9: Feature importance analysis from Random Forest for False Positive (FP) and False Negative (FN) classifications. The table reports both impurity based importance and permutation importance (accompanied by the standard deviation across 20 repetitions) for each feature. A dash (-) indicates that the value was not computed or not meaningful for that feature, typically due to low prevalence or insufficient variation in the subset.

and `hand_frac`. The permutation importance analysis also reveals moderate contributions from variables such as `n_regions` and `attenuate_at_the_base_freq`, suggesting that complex feature interactions and nonlinear relationships play a role in error formation. Overall, FP and FN patterns are broadly comparable, reflecting similar underlying feature dynamics.

4.4.2 Metric-specific correlation with accuracy

Region Coverage

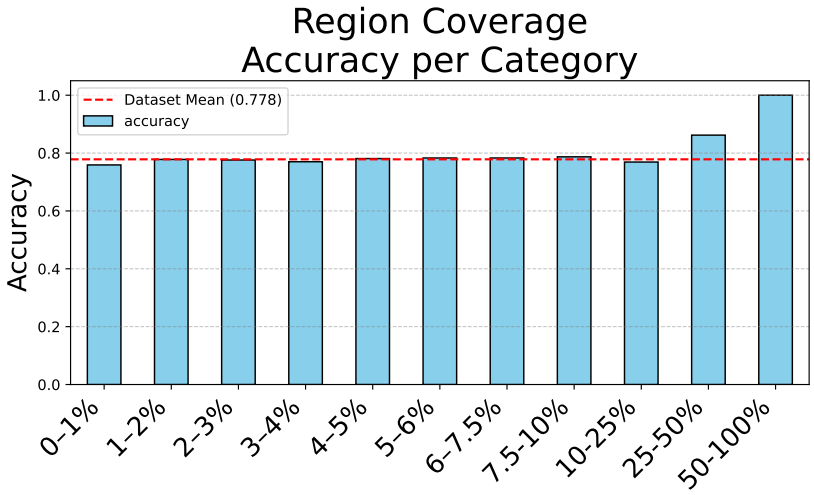
As indicated by the Random Forest analysis, the fraction of the image occupied by extracted regions (`coverage_frac`) was identified as the most influential predictor of model correctness. To explore this relationship in greater detail, we examined how prediction accuracy changes across different levels of region coverage. The images were divided into eleven coverage categories, and for each bin we calculated the mean accuracy, the number of samples, and the corresponding species distribution divergence using the KL divergence metric (Tab. 4.10, Fig. 4.18).

Category	Accuracy	Sample size	KL Divergence
0-1%	0.759	220	0.009
1-2%	0.778	2977	0.007
2-3%	0.776	2134	0.008
3-4%	0.770	3891	0.005
4-5%	0.781	3419	0.003
5-6%	0.783	3754	0.032
6-7.5%	0.783	5086	0.002
7.5-10%	0.787	6035	0.003
10-25%	0.769	6807	0.007
25-50%	0.862	65	0.201
50-100%	1.000	1	1.22

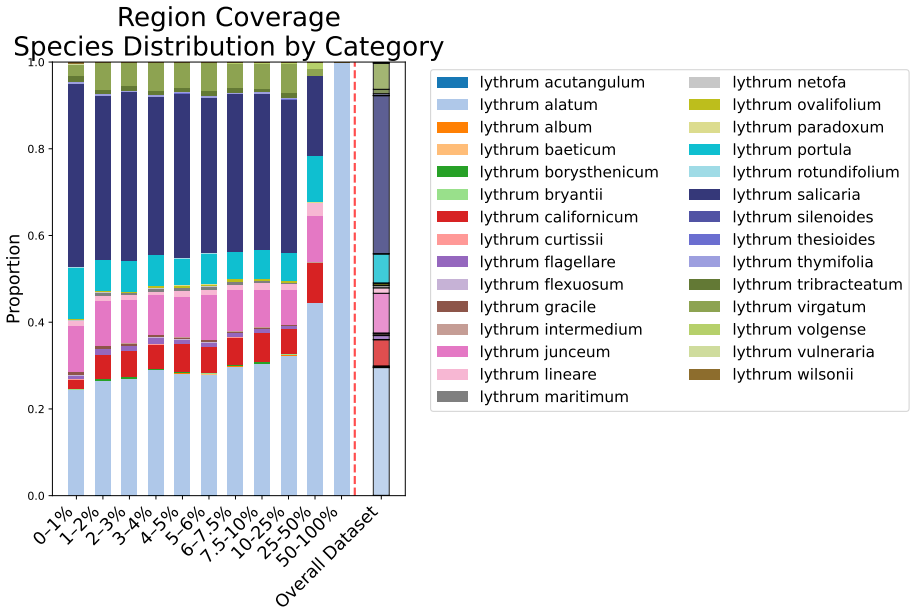
Table 4.10: Results of Region coverage correlation with accuracy analysis. For each region coverage category, mean accuracy for the category, sample size and KL divergence from the dataset distribution is shown.

ANOVA: $F = 1.221$, $p = 0.2712$; the highest mean accuracy is observed in the 50-100% coverage category (1.000) and the lowest in the 0-1% category (0.759). No strong statistical evidence of differences across categories.

Overall, the results show that classification accuracy remains largely consistent across most coverage ranges, varying between 0.76 and 0.79 up to 25% coverage (Fig. 4.18a). Only the lowest ($< 1\%$) and highest ($> 25\%$) coverage bins display noticeable deviations, although these categories contain very few samples (220 and 65 images, respectively), which limits the reliability of their estimates. The exceptionally high accuracy observed in the final bin (100%) is based on a single image and is therefore not meaningful. The species distribution divergence (KL) remains low across all main bins (< 0.01), confirming that the accuracy trends are not the result of taxonomic composition differences (Fig. 4.18b).



(a)



(b)

Figure 4.18: For the Region coverage analysis, accuracy for each category (a) and distribution of species for each category (b) is shown.

These findings indicate that, despite the high importance assigned by the Random Forest model, region coverage alone does not strongly determine classification accuracy. The Random Forest may have interpreted the separation between low and high-coverage samples as a discriminative signal, even though the underlying

relationship is weak or non-causal. This interpretation aligns with how impurity-based feature importance functions: features that allow a clear division of the dataset, even for a limited subset of samples, can receive disproportionately high importance values. In this case, a small number of images with very high coverage also exhibit higher accuracy and are easily distinguishable from the rest. As a result, the model identifies coverage as an effective split criterion, even though its predictive relevance across the dataset is limited.

It is also worth noting that the vast majority of images in the dataset display very low coverage, while high-coverage samples are comparatively rare and show larger KL divergence values with respect to the overall species distribution. This suggests that such images are not representative of the dataset as a whole and may bias the Random Forest's assessment of feature importance. Consequently, within the available data, there is **no clear evidence of a consistent relationship between coverage and prediction accuracy**, although a potential link cannot be completely excluded due to the strong imbalance in coverage distribution.

Beyond region coverage, the same analysis was extended to other metrics identified as potentially relevant by the Random Forest model, including the Pielou evenness index, the number of distinct traits (richness), hand and background fractions, and image complexity. The detailed results of these analyses are provided in Appendix C. In general, no other metric demonstrated a clear or consistent association with prediction accuracy, aside from weak or dataset-dependent effects. While some metrics displayed statistically significant trends, these were often accompanied by large differences in species composition across bins (high KL divergence), suggesting that the observed patterns are more likely driven by taxonomic imbalance than by intrinsic effects of the metrics themselves.

4.4.3 Pairwise Trait Importance and Masked Image Analysis

In Sec. 3.5.3, we introduced a modified Random Forest analysis that differs from the previous one by using combinations of trait pairs as input features. In this section, we present the corresponding results. The analysis produced feature importance scores for each trait pair, calculated both globally and separately for each prediction category (TP, TN, FP, FN). Tab. 4.11 reports the top 15 pairs and their respective importance values across these categories. The results were stable across multiple Random Forest runs, showing only small variations in the precise ranking of pairs.

Based on the results, we selected five pairs of traits for detailed examination (Tab. 4.12). These pairs were not chosen solely according to their global importance but rather for their relevance within the True Positive and True Negative categories. This criterion ensured that the selected combinations were most informative for correctly identifying both invasive and non-invasive species. The chosen pairs

Feature A	Feature B	Importance				
		Global	TP	TN	FP	FN
erect	sessile	0.3112	0.3051	0.3005	0.3055	0.3091
opposite	sessile	0.1811	0.1755	0.1823	0.1326	0.1908
alternate	subsessile	0.1777	0.1945	0.1860	0.1026	0.1846
erect	opposite	0.1675	0.1508	0.1668	0.2016	0.1522
linear	opposite	0.0574	0.0733	0.0610	0.0218	0.0672
erect	linear	0.0521	0.0532	0.0554	0.0185	0.0518
opposite	petiolated	0.0267	0.0217	0.0235	0.0716	0.0202
opposite	subsessile	0.0065	0.0065	0.0064	0.0095	0.0063
opposite	prostrate	0.0041	0.0044	0.0041	0.0040	0.0038
obovate	sessile	0.0027	0.0026	0.0026	0.0038	0.0023
prostrate	subsessile	0.0025	0.0024	0.0025	0.0034	0.0025
alternate	sessile	0.0024	0.0024	0.0017	0.0769	0.0022
alternate	erect	0.0019	0.0016	0.0017	0.0144	0.0014
erect	obovate	0.0018	0.0031	0.0020	0.0026	0.0029
alternate	creeping	0.0017	0.0009	0.0012	0.0126	0.0009

Table 4.11: Feature pair importances across global and per error-type outcomes (TP, TN, FP, FN) for top 15 pairs.

belong either to the *Common* category, shared by invasive and non-invasive taxa, or to the *Non-Invasive only* category, which includes traits specific to non-invasive species. No pairs from the *Invasive only* category were present in the dataset.

Feature A		Feature B		Category
Name	Plant Structure	Name	Plant Structure	
Erect	Stem	Sessile	Leaf	Common
Alternate	Leaf	Subsessile	Leaf	Non-Invasive only
Opposite	Leaf	Sessile	Leaf	Common
Erect	Stem	Opposite	Leaf	Common
Linear	Leaf	Opposite	Leaf	Non-Invasive only

Table 4.12: Selected pairs of traits for the pairwise analysis and their corresponding category. For each trait is also specified the plant structure which refers to.

For each selected pair of traits, we identified all images containing both traits and located the corresponding regions within those images that exhibited at least one of them. The counts for these occurrences are summarized in Tab. 4.13. This subset of images and regions represents the data targeted for removal in the subsequent masked-image experiments. Examples of these images, along with the highlighted

trait regions, are shown in Fig.4.19 and Fig.4.20.

Figure 4.19: One of the images which contained the characteristic traits **Linear-Opposite**, representing a *Lythrum californicum*. Both the original image and the image with the masked regions were classified as *Invasive* despite being *Non-Invasive*. However, after masking the regions containing the traits into consideration, the classifier was 17.1% more confident into predicting the image as *Invasive* (82.1% vs 100%).

Pairs of Traits	**Number of Images**	**Number of Regions**	**Avg Region per Image**
Erect - Sessile	19844	49674	2.50
Alternate - Subsessile (NI)	2693	6011	2.23
Opposite - Sessile	24083	55310	2.30
Erect - Opposite	21567	53982	2.50
Linear - Opposite (NI)	1857	4311	2.32

Table 4.13: For each selected pair of traits, the table shows the number of images containing both traits, the total number of regions within those images that include at least one trait (of the pair) and the average number of considered region per image. Pairs belonging to the *Non-Invasive only* category are tagged with (NI), while the remaining pairs have the *Common* category.

To evaluate the effect of removing the most predictive trait combinations on model behavior, we measured both classification performance and prediction confidence using the masked images. Tab. 4.14 summarizes the results in terms of overall accuracy, the number and proportion of label flips (images whose predicted label changed after masking), and the variation in True Positive (TP) and True

Figure 4.20: One of the images which contained the characteristic traits **Erect-Opposite**, representing a *Lythrum virgatum*. The original image was correctly classified as *Invasive* with 77.0% confidence in the prediction. After masking the regions containing one or more traits into consideration, the classifier predicted the image to be *Non-Invasive* with 100% confidence in the prediction.

Negative (TN) counts. The corresponding accuracy changes for each analyzed trait pair are illustrated in Fig. 4.21.

Pairs of Traits	Accuracy			Flips		TP counts			TN counts		
	Old	New	Δ	Count	Rate	Old	New	Δ	Old	New	Δ
Erect - Sessile	0.709	0.707	-0.002	3426	17.3%	6836	6756	-80	7239	7281	+42
Alternate - Subsessile (NI)	0.973	0.268	-0.705	1920	71.3%		-		2619	721	-1898
Opposite - Sessile	0.733	0.713	-0.020	4242	17.6%	7530	7437	-93	10120	9743	-377
Erect - Opposite	0.729	0.722	-0.007	3593	16.7%	6836	6756	-80	8896	8821	-75
Linear - Opposite (NI)	0.959	0.889	-0.070	187	10.1%		-		1781	1652	-129

Table 4.14: Results of the masked images analysis. For each considered pair of traits is shown: accuracy before (old) and after (new) images were masked (with relative difference (Δ computed), the number and the rate of flips of prediction (i.e., times when the prediction of model changes), True Positive (TP) and True Negative (TN) counts before (old) and after (new) images were masked (with relative difference (Δ computed). Pairs belonging to the *Non-Invasive only* category are tagged with (NI), while the remaining pairs have the *Common* category.

For the pairs belonging to the *Common* category, the effect of masking was limited. The overall accuracy dropped by less than 2%, and the rate of label flips remained close to 17%. Changes in TP and TN counts were also minimal,

suggesting that these trait combinations, although frequently observed, are not essential for the classifier to correctly predict plant invasiveness.

In contrast, masking pairs from the *Non-Invasive only* category produced much stronger effects. In particular, removing regions corresponding to the pair *Alternate-Subsessile* resulted in a substantial decrease in accuracy (from 0.97 to 0.27) and caused more than 70% of the images to change their predicted class. Similarly, masking the pair *Linear-Opposite*, although less dramatic, led to a 7% reduction in accuracy.

Pairs of Traits	**True Invasive**			**True Non-Invasive**		
	n images	**Mean** $\Delta P(Invasive)$	**SD** ΔP	**n images**	**Mean** $\Delta P(Invasive)$	**SD** ΔP
Erect - Sessile	11357	+0.023	0.387	8487	-0.021	0.286
Alternate - Subsessile (NI)		-		2693	+0.694	0.448
Opposite - Sessile	12545	+0.022	0.388	11538	+0.015	0.320
Erect - Opposite	11357	+0.023	0.387	10210	-0.010	0.288
Linear - Opposite (NI)		-		1857	+0.046	0.293

Table 4.15: Changes in predicted probabilities in the masked images analysis. For each considered pair of traits, images were divided according to their true class (Invasive or Non-Invasive). For each subset, the table reports the number of images (n images), the mean change in the predicted probability of the *Invasive* class ($\Delta P(Invasive)$), and its standard deviation. Pairs belonging to the *Non-Invasive only* category are tagged with (NI), while the remaining pairs have the *Common* category.

For the pairs belonging to the *Common* category (*Erect-Sessile*, *Opposite-Sessile*, and *Erect-Opposite*), the effect of masking on model confidence was limited, as shown in Tab. 4.15 and Fig. 4.22. On average, masking these traits resulted in only minor variations in the predicted probability of the true class (mean $\Delta P(true\ class) \approx +0.02$), confirming that these trait combinations were not decisive factors in the model's final predictions.

Across all three pairs, the distributions of $\Delta P(true\ class)$ were centered near zero for both true invasive and true non-invasive images, indicating a balanced and overall weak effect. The confidence bin analysis (Fig. 4.22b) shows that for true invasive samples, masking slightly reduced the predicted probability of the invasive class at high confidence levels (0.8−1.0 bin), while for low-confidence images (0−0.4 bins) small positive shifts were occasionally observed. A similar trend appeared for true non-invasive samples, with slight positive shifts at low confidence and mild negative ones at high confidence, suggesting that the model remained stable after trait removal.

These findings indicate that *Common* trait pairs represent broadly informative morphological patterns shared across taxa, but do not act as decisive cues for

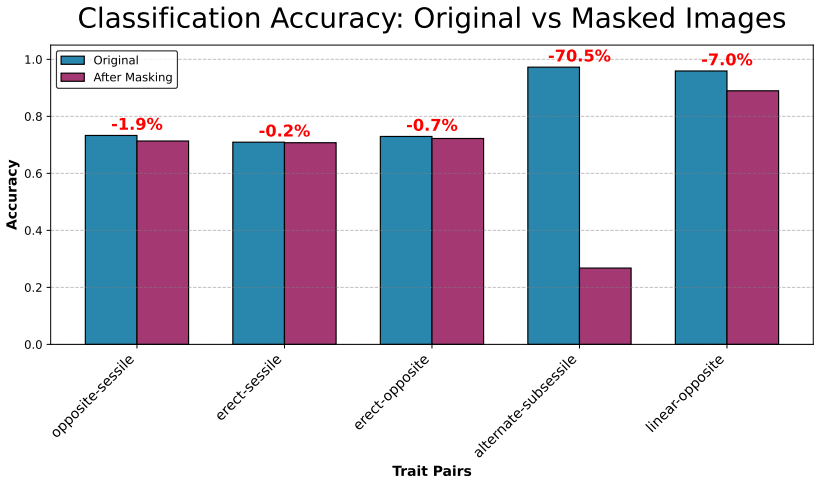


Figure 4.21: For each pair of traits, accuracy computed with original images (in blue) and accuracy computed with masked images (in magenta) are shown; on top of each bar, the difference in accuracy is computed in red.

differentiating invasive from non-invasive plants. Their removal produces only small redistributions of prediction confidence without substantial effects on accuracy or classification direction. This modest impact may also reflect that in many images, not all occurrences of the traits were removed, allowing the model to rely on the remaining ones to maintain stable predictions.

In contrast, the pairs belonging to the *Non-Invasive only* category produced a much stronger and more directional response to masking, as illustrated in Tab. 4.15 and Fig. 4.23. When regions associated with *Alternate-Subsessile* were removed, the predicted probability for the non-invasive class decreased markedly (mean $\Delta P(non\ invasive\ class) = -0.69$), corresponding to a significant drop in accuracy. As shown in Fig. 4.23b, this effect became more pronounced at higher confidence levels: for samples originally classified as confidently non-invasive (confidence bins > 0.6), masking consistently produced large negative shifts in ΔP, indicating that the model became less confident or even reversed its prediction after trait removal. This pattern suggests that *Alternate-Subsessile* serves as a strong and distinctive visual cue for non-invasiveness, whose absence leads the classifier to favor invasive predictions.

The second pair, *Linear-Opposite*, exhibited a similar trend but with lower intensity (mean $\Delta P(non\ invasive\ class) = -0.05$). The per-bin analysis shows small positive shifts at low confidence and slight negative shifts at high confidence, implying a modest but consistent role in reinforcing non-invasive predictions. Compared with *Alternate-Subsessile*, this pair appears less diagnostic but still

contributes to maintaining model confidence for correctly classified non-invasive images.

Both *Non-Invasive only* trait pairs generated unidirectional effects on model confidence, confirming that the classifier relied on these visual combinations as characteristic indicators of non-invasive species. Their removal systematically biased the predictions toward the invasive class, emphasizing their relevance as discriminative, class-specific features.

In summary, these results demonstrate that masking common trait combinations has only a minor impact on predictions, likely because they encode redundant visual information, whereas removing traits distinctive of non-invasive species substantially affects both classification accuracy and confidence, highlighting their importance in the decision process.

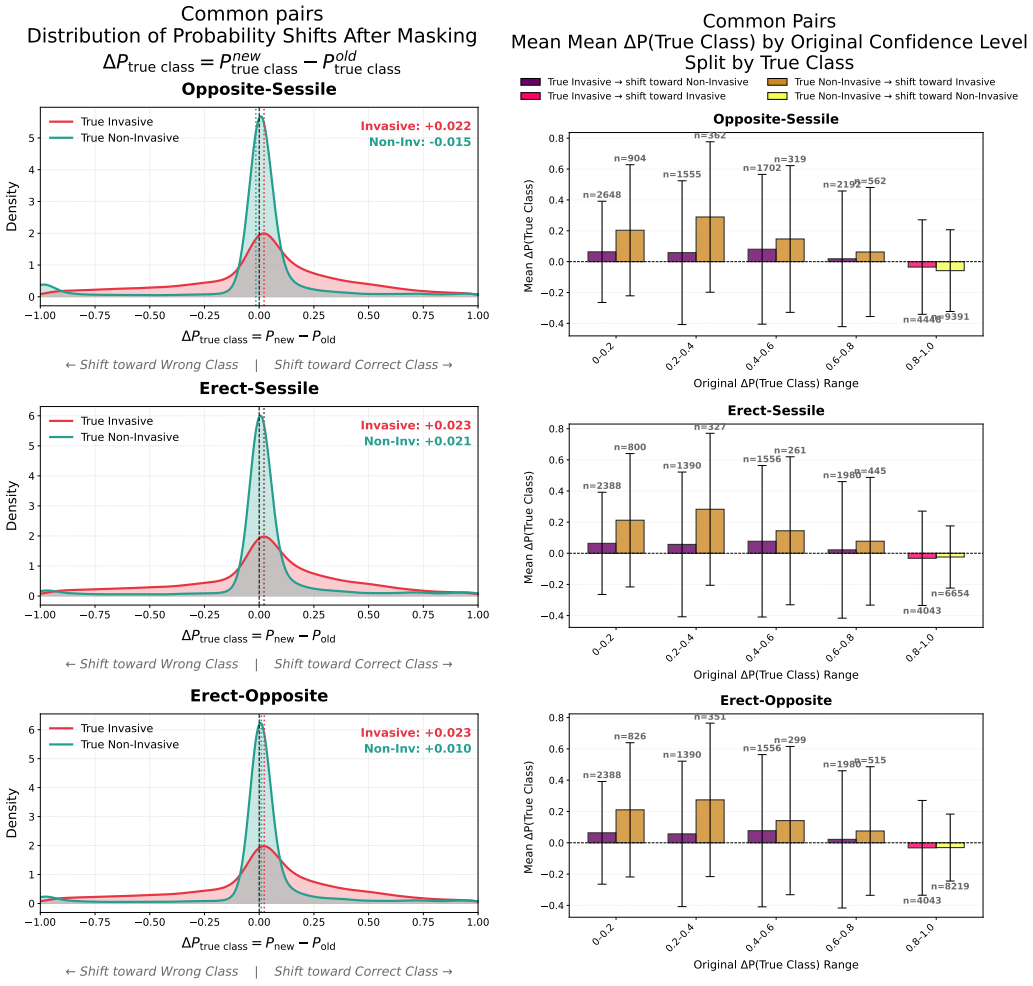


Figure 4.22: Effect of masking on model prediction probabilities for each trait pair belonging to the *Common* category. (a) Distribution of changes in the predicted probability of the true class ($\Delta P(true\ class) = P^{new}_{true\ class} - P^{old}_{true\ class}$) for true Invasive and true Non-Invasive images. Positive values indicate a shift toward the correct class after masking, while negative values indicate a shift toward the wrong class. Mean $\Delta P(Invasive)$ values for each class are annotated in the top-right corner of each subplot. (b) Mean change in predicted probability ($\Delta P(true\ class)$) across bins of the model's original confidence, split by true class. Bars represent mean shifts with error bars denoting standard deviation. For *True Invasive* images, dark purple bars indicate shift toward the Invasive class and bright pink bars indicate shifts toward the Non Invasive class. For *True Non Invasive* images, dark orange bars indicate shifts toward the Non Invasive class and light yellow bars indicate shifts toward the Invasive class.

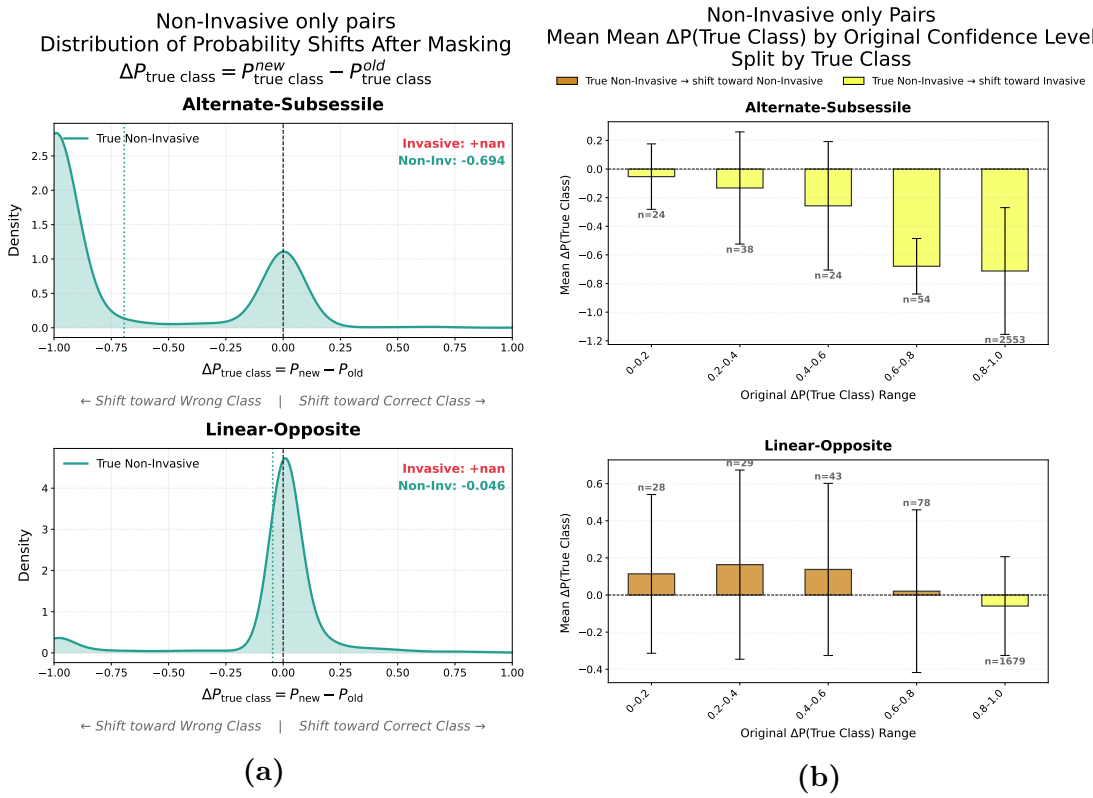


Figure 4.23: Effect of masking on model prediction probabilities for each trait pair belonging to the *Non Invasive only* category. (a) Distribution of changes in the predicted probability of the true class ($\Delta P(true\ class) = P^{new}_{true\ class} - P^{old}_{true\ class}$) for true Invasive and true Non-Invasive images. Positive values indicate a shift toward the correct class after masking, while negative values indicate a shift toward the wrong class. Mean $\Delta P(Invasive)$ values for each class are annotated in the top-right corner of each subplot. (b) Mean change in predicted probability ($\Delta P(true\ class)$) across bins of the model's original confidence, split by true class. Bars represent mean shifts with error bars denoting standard deviation. For *True Invasive* images, dark purple bars indicate shift toward the Invasive class and bright pink bars indicate shifts toward the Non Invasive class. For *True Non Invasive* images, dark orange bars indicate shifts toward the Non Invasive class and light yellow bars indicate shifts toward the Invasive class.

Chapter 5

Conclusions

In these thesis we have presented a pipeline to identify morphological traits related to the potential invasiveness of a species of plants inside its genus, relying on image data only. We used the *Lythrum* genus as a case study, aiming to identify which characteristic traits were important to predict if a certain species was invasive or non-invasive.

First, we retrieved our image data from `iNaturalist.org`, a platform where users can upload pictures of living organisms, making them available for citizen science projects. Then we employed the state-of-the-art computer vision model BioCLIP 2 to embed the images in our dataset: BioCLIP 2 proved to be suitable for the task, differentiating species with minimal morphological differences, and providing fine-grained representation of the data. A classifier was trained using these embeddings, with the goal of predicting whether each input image represented an *Invasive* or a *Non-Invasive* species. In particular, one classifier per species was utilized: it was trained on every image except for the ones representing that species, so that for subsequent analysis the classifier would have never seen the species under scrutiny.

We then introduced explainability into the pipeline: we employed Integrated Gradients, an XAI algorithm able to produce feature attribution maps, to identify for each image the regions that the model considered most important for the prediction. We then clustered the regions, manually labeling them with either one or more structures (*Leaf, Flower, Stem*) or a non-informative feature (*Hand, Background/undefined*) according to what they represented.

Finally, with the extracted regions from each image we are able to mine patterns and extract information on the characteristic traits of the invasive or non-invasive plants. We do so by aggregating the regions and the corresponding labels back into the original images, computing several different metrics for each of them. We separate the predictions of the original models according to their type (True Positive, True Negative, False Positive or False Negative) and we apply a Random

Forest classifier to identify which of the aforementioned metrics or image features are more related to the prediction correctness. Struggling to obtain meaningful results when taking into consideration only one feature, we performed a *pairwise* trait importance analysis, combining the traits into couples and grouping them according to their specificity for every class. We take the top 5 pairs for feature importance from the Random Forest and we look at the prediction for the images when masking the regions that contain the traits into consideration. We find out that for the three pairs common to both invasive and non-invasive species the accuracy doesn't deviate greatly from the original results. We are satisfied to report, however, that for the two pairs that are exclusive to the non-invasive species (*Alternate-Subsessile* and *Linear-Opposite*), when masking the regions that contain them and therefore removing these traits from the image, the accuracy sees a significant drop, with the classifier changing its prediction for several images. In particular, when masking the traits relative to the pair *Alternate-Subsessile*, the images have a drop in accuracy by 70.5%, with 71.3% of the images changing their label from the original prediction to the masked prediction. This suggests, in accordance with the direction of our research, that it is possible to detect visual traits that predict non-invasiveness. Due to limitations in the dataset we were not able to detect traits correlated to invasiveness, since there are no pairs of characteristic traits that are exclusive to invasive species, but the obtained results are promising nonetheless. It is important to notice that all the top 5 pair of traits (both the 'common' ones and the 'non-invasive' ones) taken into consideration result in a drop in accuracy when masked. This is significant for the validation of our methodology, indicating that it could be adopted and expanded by researchers that will select a similar approach.

There are a few considerations to make regarding the developement of this work.

The first is that we had to remove *Lythrum hyssopifolia* from the analysis because the model was unable to classify it correctly. Although justified by the improvement of the results, this choice was dictated by the identification of *L. hyssopifolia* as an analitic outlier, not a biological one. There aren't, to the best of our knowledge, research works that identify anomalies in the invasiveness of *L. hyssopifolia*, or that might explain why the model recognize it so differently from the other two invasive species present in the dataset. Contrarily to our assumption, this might indicate that either the model is failing to embed *L. hyssopifolia* as a species, or that not all invasive species can be identified according to visual traits only. We invite future researchers that will pick up on our work to investigate further this topic.

The dataset itself presents other possible limitations: after the removal of *Lythrum hyssopifolia* only two invasive species remain, compared to twenty-seven non-invasive species (despite the sample sizes of the two class remaining only slightly unbalanced). This removes variability for the model to learn on, forcing

it to rely only on two species and their images. Future works that will follow an approach comparable to ours should try to work with a more diverse dataset that will allow the model to capture more traits and difference between species, eventually exploring other genera or families altogether.

Another limitation in the approach can be identified in the explainability pipeline: the original images were not downloaded with the highest possible quality due to limited storage capacity and API usage bottlenecks. This does not affect the performance of the model, as demonstrated by our results, but makes it so that some of the extracted regions have a low resolution, with a possible impact on their embeddings or their analysis.

The labeling of the different clusters, then, despite being validated by comparing the results of two different algorithms is an approximation: it comes from a manual analysis (performed by a small group of thesists) of a subset of regions instead of the whole dataset. The manual labeling however was necessary, as we are not aware of any pre-trained segmentation model or classification model able to identify the traits into consideration from the images.

The traits we used to label the clusters were few and generic, but they had to be that way since they allowed us to be certain of what was included in the region (e.g. a flower or stem). In future works it might be interesting to explore ways to include sub-traits (e.g. flower color or internode), to enrich the analysis and furthermore increase its relevance.

Finally, we have broadened the scope of our research by labeling every species that was invasive in any part of the world as 'Invasive' (as it is reasonable to expect interspecific differences to be more pronounced than intraspecific ones, i.e. between native and invasive population of the same species). This was because, due to the way the data was gathered, we had access to only a limited subset of geolocalized images. We made the research choice to have access to a greater amount of data, discarding the information about the image location. Future works that will take a similar task are invited to expand the research question, identifying a genus or a family that allows them to work with both geographical information and a sufficient amount of data, to investigate the role of potentially invasive species in locations where they are actually invasive, in locations where they are alien but do not become invasive, and in locations where they are instead native, to test whether invasiveness is a property of a species per se, or if it depends on the ecological context the species finds itself in.

Other than geographical information, it could be interesting to integrate other metadata (not exclusively morphological, such as phylogeny, temporal information on the pictures, metereological information on the location...) into the pipeline, to enrich the type and quantity of available information used for the analysis, and study if there is a correlation between categorical and numerical traits, and visual traits (that we took into consideration for this project).

In conclusion, in this thesis we showed how it's possible to identify invasive species using exclusively visual traits obtained from images, and what taxon-specific traits are important for the prediction. This study adds to a growing body of work investigating the morphological features that characterize invasive species, helping with their identification and mitigation of effects.

Appendix A

Labels enrichment with species characteristic traits

We enhanced the labels inherited from the cluster assignments (see Sec. 4.3.3) by incorporating specific traits defined for each species and plant structure. For every region extracted from the dataset, we first identified the corresponding species. After the clustering phase, each region was already labeled according to the biological structures it contained (*Leaf*, *Flower*, or *Stem*). Based on this information, we assigned to each region the characteristic traits associated with the identified structures for that particular species, as listed in Tab. A.1. For instance, a region of *Lythrum anatolicum* labeled as containing both a flower and a stem would be annotated with the traits *Petals purple, Stamens 12* for the flower and *Erect* for the stem.

Species	**Leaf**	**Flower**	**Stem**
Lythrum acutangulum	-	-	-
Lythrum alatum	Opposite Opposite, becoming alternate distally Sessile Attenuate at the base	Inflorescence raceme Petals purple Floral tube cylindrical Stamens 6	Erect
Lythrum album	Alternate	Petals white	Erect
Lythrum anatolicum	Opposite Sessile Cordate at the base	Petals purple Stamens 12	Erect
Lythrum baeticum	-	-	-
Lythrum borysthenicum	Opposite, becoming alternate distally Sessile Obovate	Petals reddish Petals minute Stamens 6	Erect Erect or decumbent
Lythrum bryantii	-	-	-
Lythrum californicum	Opposite Opposite, becoming alternate distally Linear	Inflorescence raceme Petals purple Stamens 5-8	Erect
Lythrum curtissii	Opposite Opposite, becoming alternate distally	Inflorescence raceme Petals lilac to pink	Erect

Species	**Leaf**	**Flower**	**Stem**
	Sessile or subsessile	Usually with a darker midrib	
	Attenuate at the base	Floral tube obconic	
		Stamens 6	
Lythrum flagellare	Opposite	Inflorescence raceme	Creeping to weakly erect
	Petiolated	Floral tube obconic without red dots	
	Rounded at the base	Petals purple	
	A discernible gap between the stem and the base of the blade	Stamens 6	
Lythrum flexuosum	Alternate	Calyx with alternate long and small teeth	Creeping
	Sessile	Flowers solitary	
		Petals purple	
Lythrum gracile	Opposite	Petals white to pink	Erect
	Opposite, becoming alternate distally		
	Rounded at the base		
Lythrum hyssopifolia	Alternate	Inflorescence raceme	Erect to weakly erect
	Sessile	Floral tube obconic without red dots	
	Rounded at the base	Petals pink	
		Calyx with alternate long and small teeth	
		Stamens 4-6	
Lythrum intermedium	Opposite	Inflorescence spikelike	Erect
	Sessile	Flowers in whorled clusters	
	Opposite, becoming alternate distally	Calyx with alternate long and small teeth	
	Rounded at the base	Floral tube cylindrical	
		Petals purple	
		Stamens 12	
Lythrum junceum	Alternate	Inflorescence raceme	Sprawling or ascending
	Subsessile	Flowers solitary in leaf axils	
	Obtuse to truncate at the base	Floral tube obconic	
		Floral tube red dotted	
		Petals purple	
		Stamens 12	
Lythrum lineare	Opposite	Inflorescence raceme	Erect
	Sessile	Floral tube cylindrical	
	Attenuate at the base	Petals pale purple or whitish	
		Usually with a darker midrib	

Species	**Leaf**	**Flower**	**Stem**
		Stamens 6	
Lythrum maritimum	Opposite Subsessile	Petals pink Usually with a darker midrib	Prostrate
Lythrum netofa	Sessile Broader at the base	Flowers solitary Floral tube campanulate Petals 4 Petals purple with a darker midrib Stamens 6-8	Erect
Lythrum ovalifolium	Alternate Sessile or subsessile Attenuate at the base	Inflorescence raceme Floral tube obconic without red dots Petals purple with a darker midrib Stamens 6	Erect or decumbent
Lythrum paradoxum	Alternate Sessile	Petals pink to purple Stamens 10-12	Erect
Lythrum portula	Opposite Sessile	Inflorescence spikelike Floral tube campanulate Petals white to pink Stamens 5-8	Prostrate and spreading
Lythrum rotundifolium	Petiolated	Flowers solitary Petals pink to purple with a darker midrib Stamens 8	Prostrate
Lythrum salicaria	Opposite Sessile Opposite, becoming alternate distally Rounded at the base	Inflorescence spikelike Flowers in whorled clusters Calyx with alternate long and small teeth Floral tube cylindrical Petals purple Stamens 12	Erect
Lythrum silenoides	-	-	-
Lythrum thesioides	Alternate	Petals 4 Petals pink	Erect
Lythrum thymifolia	Needle-like	One or few flowers in the axil of leaves Calyx with alternate long and small teeth Stamens 2-3	Prostrate
Lythrum tribracteatum	Opposite Sessile Attenuate at the base	Inflorescence spikelike Floral tube narrowly cylindrical without red dots Calyx with teeth of the same length	Prostrate to weakly erect

Species	**Leaf**	**Flower**	**Stem**
		Petals lavender Stamens 4-6	
Lythrum virgatum	Opposite Sessile Narrower at the base	Inflorescence spikelike to raceme Flowers in whorled clusters Floral tube cylindrical Petals pink to purple Stamens 10-14	Erect
Lythrum volgense	Needle-like	Petals pink Petals minute	Prostrate
Lythrum vulneraria	Opposite Opposite, becoming alternate distally	Petals pink	Erect
Lythrum wilsonii	Alternate Sessile Rounded at the base	Petals pink to purple	Erect

Table A.1: Characteristic traits for each species for each biological structure.

Appendix B

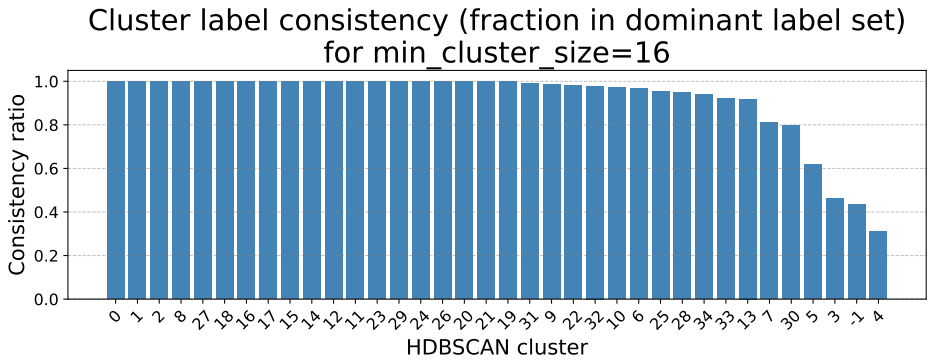
HDBSCAN clustering validation details

In addition to the summary presented in the main text (see Sec. 4.3.3), we examined cluster consistency and label distributions for all valid HDBSCAN configurations in detail.

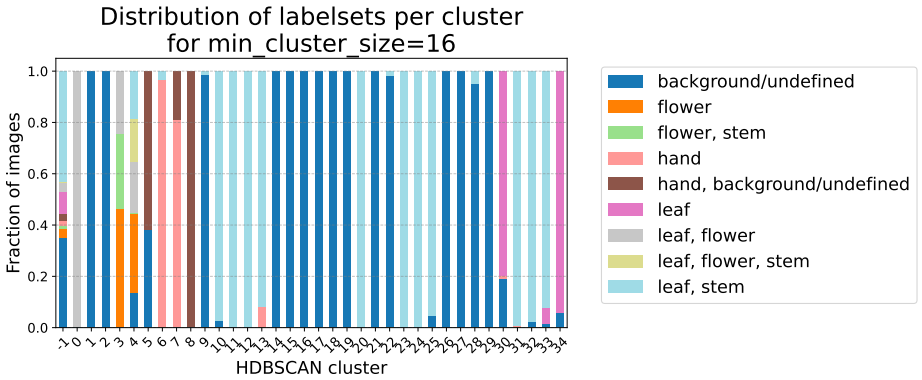
Figs. B.1 and B.2 present the per-cluster validation results for the best configuration ($min\_cluster\_size = 16$) and the worst configuration ($min\_cluster\_size = 27$), respectively. Cluster -1 corresponds to the noise cluster and is therefore excluded from the analysis. The distributions of consistency ratios (Figs. B.1a and B.2a) are consistent with the global distribution shown in Fig. 4.16a, as expected.

Figs. B.1b and B.2b illustrate the detailed distribution of label sets within each cluster. In most cases, even when multiple label sets are present within a cluster, they tend to be closely related. For example, in Fig. B.2b, cluster 2 contains the label sets *'flower'*, *'flower, leaf'*, and *'flower, stem'*, which all share common biological features. Occasionally, clusters include a mixture of biological and non-biological labels: for instance, cluster 30 in Fig. B.1b includes both *'background/undefined'* and *'leaf'*, but typically one of these label types represents only a small portion of the cluster's samples.

As expected, the configuration with $min\_cluster\_size = 27$ yields slightly poorer results. This setting produces only nine clusters (with the addition of the noise cluster bringing the total just above the threshold). As shown in Fig. B.3, the number of clusters generated by a configuration has a marked influence on the overall average consistency ratio. Fewer clusters tend to merge heterogeneous samples, reducing consistency, while a larger number of clusters enables the model to capture more homogeneous groupings.

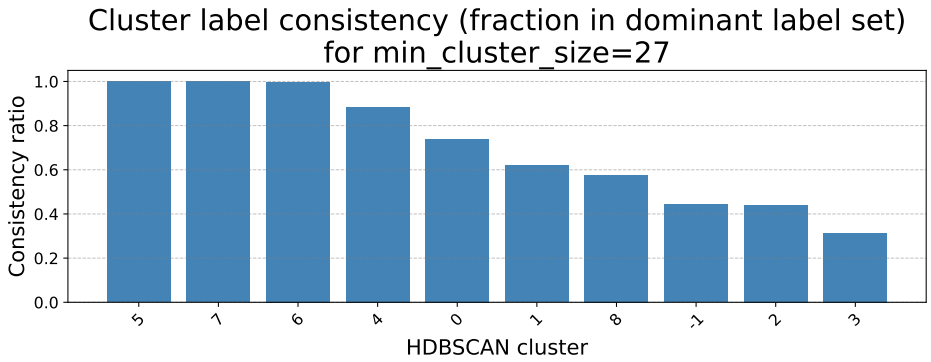


(a)

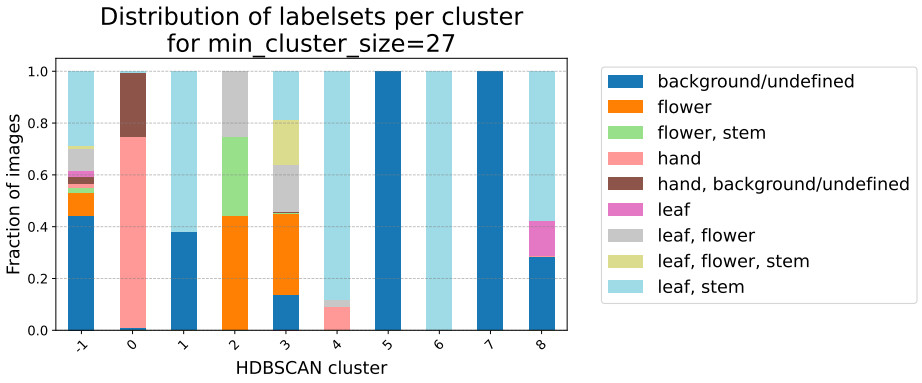


(b)

Figure B.1: Detailed clustering and labeling validation using HDBSCAN for the best configuration found. Cluster label consistency ration for each cluster (a) and distribution of labelsets per cluster (b) are shown.



(a)



(b)

Figure B.2: Detailed clustering and labeling validation using HDBSCAN for the worst configuration found. Cluster label consistency ration for each cluster (a) and distribution of labelsets per cluster (b) are shown.

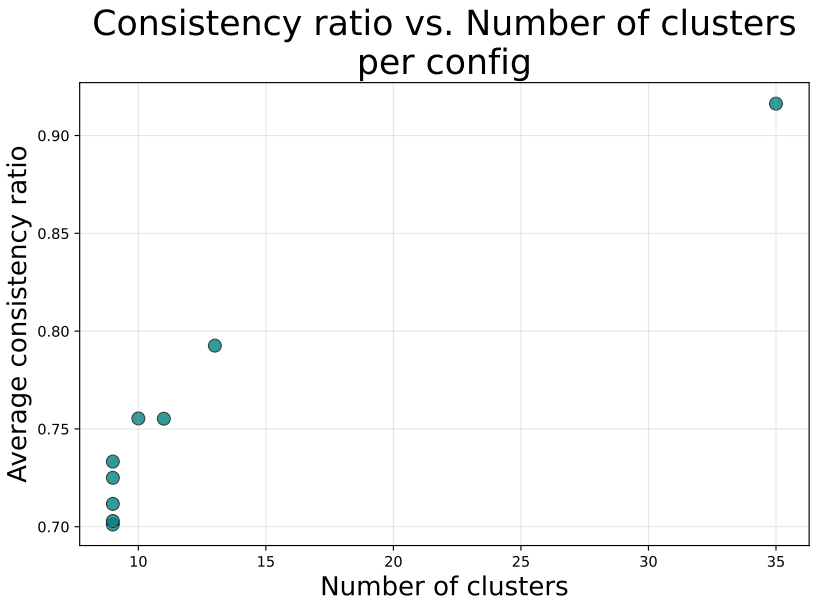


Figure B.3: Correlation between the average consistency ratio of each configuration and the number of clusters generated by the configuration.

Appendix C

Metric-specific correlation with accuracy

This appendix presents the detailed results of the metric-specific accuracy correlation analysis described in Sec. 3.4.1. For each evaluated metric, images were grouped into discrete bins, and the mean prediction accuracy, number of samples, and Kullback-Leibler (KL) divergence from the global species distribution were calculated. The objective of this analysis was to determine whether certain image-level characteristics were systematically associated with classification performance, while also assessing whether the observed effects could instead be explained by variations in taxonomic composition rather than by the metrics themselves.

C.1 Pielou Evenness

Accuracy varies across the categories of the Pielou evenness index, with a statistically significant ANOVA result ($p < 0.05$), as reported in Tab. C.1 and Fig. C.1. Images with *Simple* or *Low* evenness values (that is, those dominated by a small number of traits) exhibit slightly higher accuracy, reaching up to 0.81, compared to images with *Medium High* or *Very High* evenness, where accuracy decreases to approximately 0.74.

However, the KL divergence increases noticeably for intermediate categories, suggesting that part of the observed variation in accuracy may be due to changes in species composition rather than the evenness index itself.

Overall, the pattern suggests that images with more homogeneous trait distributions tend to be classified more accurately, although the effect remains relatively modest.

Category	Accuracy	Sample size	KL Divergence
Simple	0.807	5736	0.023
Low	0.790	5729	0.012
Medium	0.798	5734	0.021
Medium High	0.738	5764	0.056
High	0.782	7773	0.004
Very High	0.740	3653	0.060

Table C.1: Results of Pielou Evenness index correlation with accuracy analysis. For each region coverage category we report mean accuracy, sample size and KL divergence from the dataset distribution.

ANOVA: $F = 25.0784$, $p \approx 0.00$; the highest mean accuracy is observed in the 'Simple' category (0.807) and the lowest in the ''Medium High' category (0.738). Accuracy differences are statistically significant ($p < 0.05$).

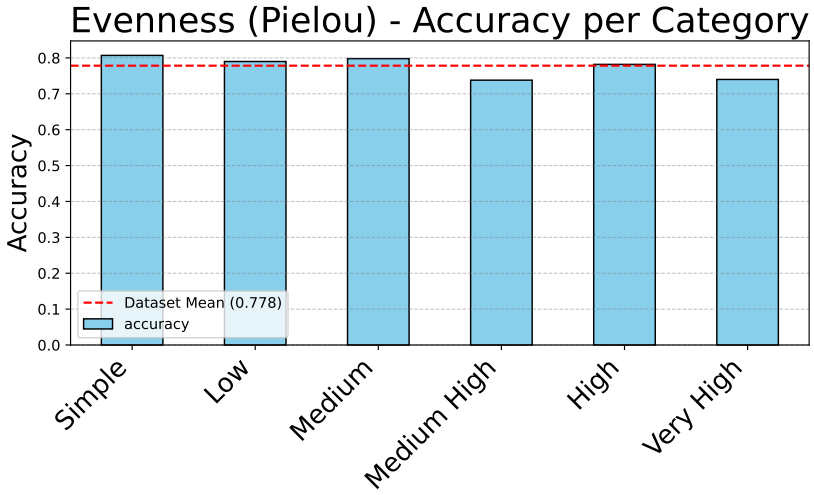
C.2 Distinct Traits (richness)

Category	Accuracy	Sample size	KL Divergence
Simple	0.790	2574	0.041
Low	0.885	3963	0.620
Medium	0.762	10390	0.072
Medium High	0.854	3954	0.580
High	0.764	11057	0.141
Very High	0.609	2451	0.759

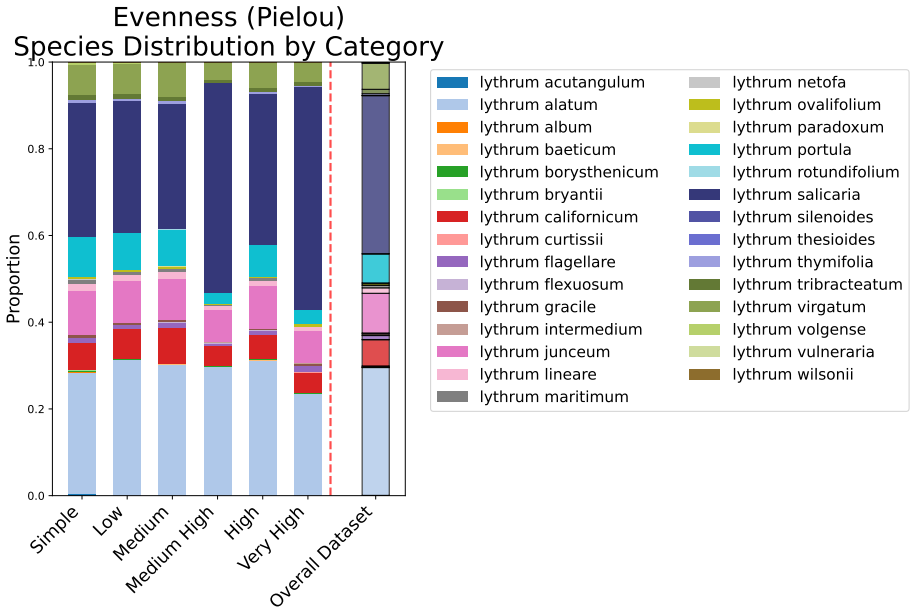
Table C.2: Results of richness (i.e., the number of distinct traits for each image) correlation with accuracy analysis. For each region coverage category we report mean accuracy, sample size and KL divergence from the dataset distribution.

ANOVA: $F = 170.03$, $p \approx 0.00$; the highest mean accuracy is observed in the 'Low' category (0.885) and the lowest in the 'Very High' category (0.609). Accuracy differences are statistically significant ($p < 0.05$).

Richness shows a statistically significant relationship with prediction accuracy ($p < 0.05$), although the pattern is irregular and clearly affected by species imbalance (see Tab. C.2 and Fig. C.2). The *Low* and *Medium High* richness categories achieve the highest accuracies (approximately 0.85–0.88), whereas both *Simple* and *Very High* richness values are associated with lower accuracies (ranging from 0.61 to 0.76).



(a)

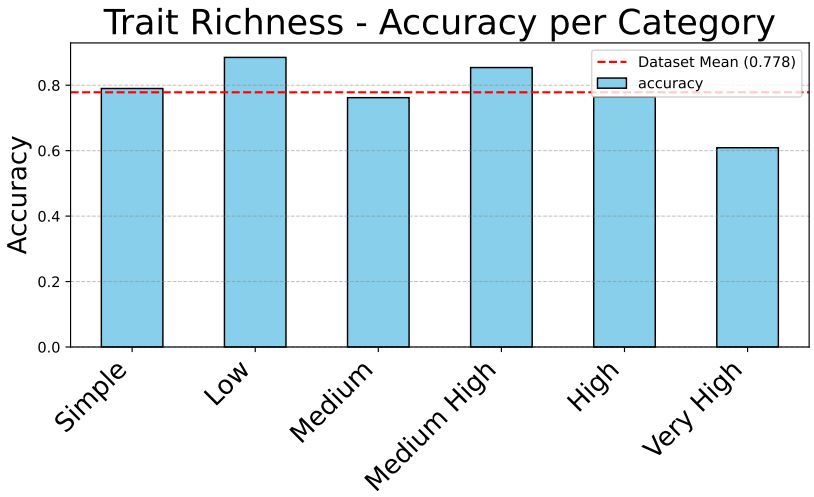


(b)

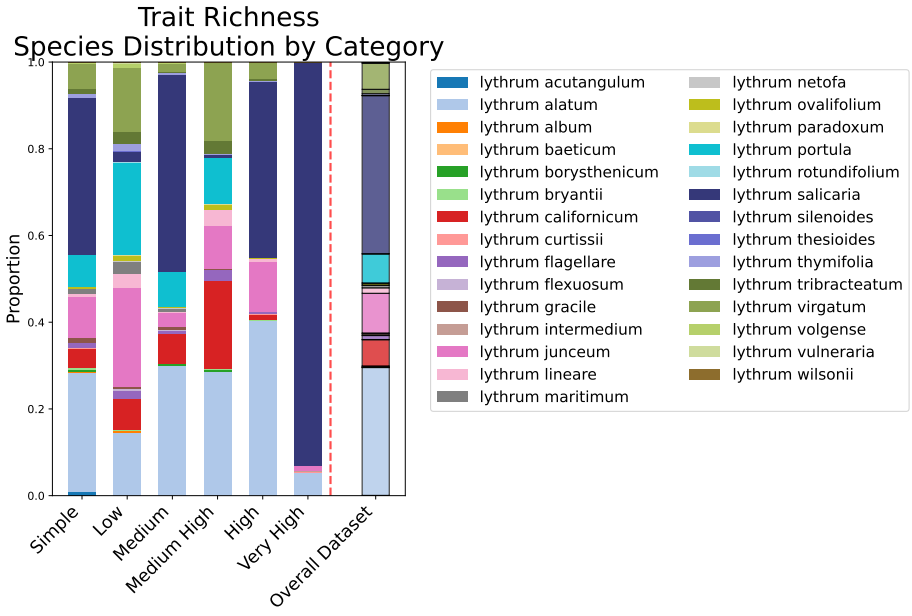
Figure C.1: For the Pielou Evenness analysis we report accuracy for each category (a) and distribution of species for each category (b).

Nevertheless, the corresponding KL divergence values are substantial (up to 0.76), indicating that bins with extreme richness levels are heavily influenced by specific taxa.

Therefore, while images exhibiting moderate trait diversity tend to yield more



(a)



(b)

Figure C.2: For the Richness analysis we report accuracy for each category (a) and distribution of species for each category (b).

stable classification results, this relationship should not be interpreted as causal.

C.3 Hand Fraction

Category	Accuracy	Sample size	KL Divergence
Simple	0.778	26988	0.001
Low	0.779	5265	0.003
Medium	0.781	1554	0.007
Medium High	0.781	187	0.071
High	0.737	19	0.144
Very High	0.763	376	0.026

Table C.3: Results of Hand fraction correlation with accuracy analysis. For each region coverage category we report mean accuracy, sample size and KL divergence from the dataset distribution.

ANOVA: $F = 0.148$, $p = 0.98$; the highest mean accuracy is observed in the 'Medium' category (0.781) and the lowest in the 'High' category (0.737). No strong statistical evidence of differences.

Accuracy remains highly consistent across all categories of hand fraction, ranging from 0.77 to 0.78, with no statistically significant differences detected (see Tab. C.3 and Fig. C.3). KL divergence values are uniformly low across all bins.

These findings suggest that the presence of human hands in the images does not systematically influence or bias the model's predictions.

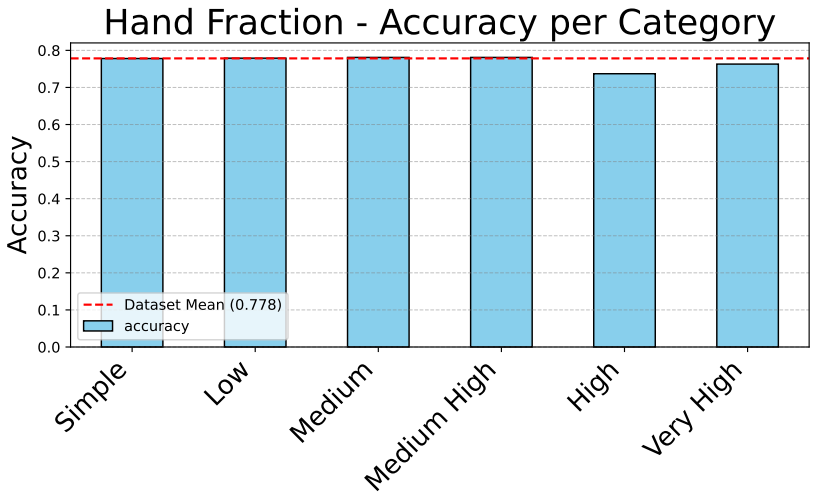
C.4 Background/undefined fraction

Similarly, Tab. C.4 and Fig. C.4 show that the fraction of regions labeled as background or undefined has no consistent association with classification accuracy. Accuracy values vary only slightly, between 0.76 and 0.79 across categories, while KL divergence remains below 0.3.

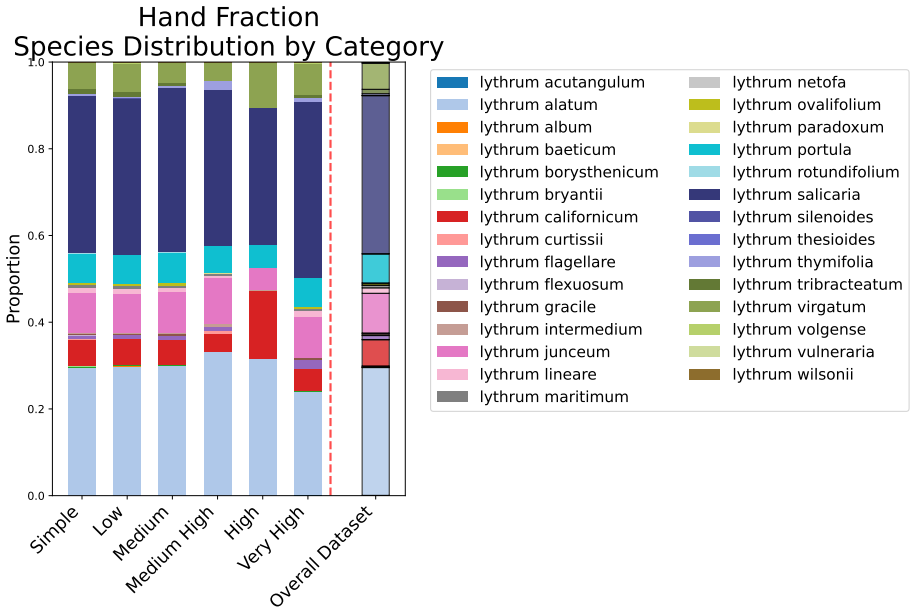
These results confirm that differences in background area do not affect model correctness, indicating that the classifier effectively concentrates on the relevant plant regions.

C.5 Image Complexity

Image complexity, expressed as the number of extracted regions per image, shows a weak but statistically significant tren (see Tab. C.5 and Fig. C.5b). Accuracy tends to increase slighlty with complexity, reaching its highest value (0.85) for very



(a)



(b)

Figure C.3: For the Hand fraction analysis we report accuracy for each category (a) and distribution of species for each category (b).

complex images (more than ten regions). Nonetheless, these categories include few samples and show higher KL divergence, indicating that the apparent trend may be driven by species composition.

Overall, while more structurally complex images might provide richer information

Category	Accuracy	Sample size	KL Divergence
Simple	0.774	13157	0.001
Low	0.780	9601	0.001
Medium	0.781	6889	0.002
Medium High	0.786	2242	0.008
High	0.761	627	0.026
Very High	0.784	1873	0.009

Table C.4: Results of Background/Undefined fraction correlation with accuracy analysis. For each region coverage category we report mean accuracy, sample size and KL divergence from the dataset distribution.

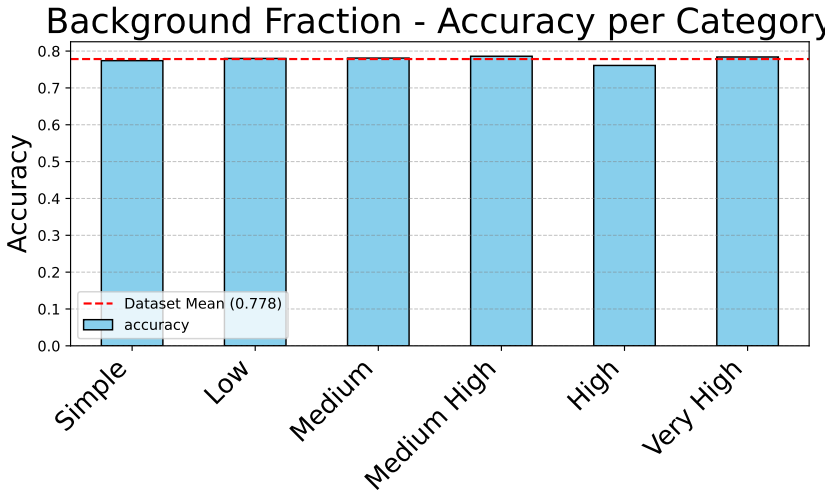
ANOVA: $F = 0.807$, $p = 0.55$; the highest mean accuracy is observed in the 'Medium High' category (0.786) and the lowest in the 'High' category (0.761). No strong statistical evidence of differences.

Category	Accuracy	Sample size	KL Divergence
Simple (1)	0.771	3978	0.008
Low (2-3)	0.775	12968	0.002
Medium (4-5)	0.786	10680	0.001
Medium High (6-7)	0.777	4820	0.007
High (8-10)	0.771	1729	0.026
Very High (10+)	0.850	214	0.098

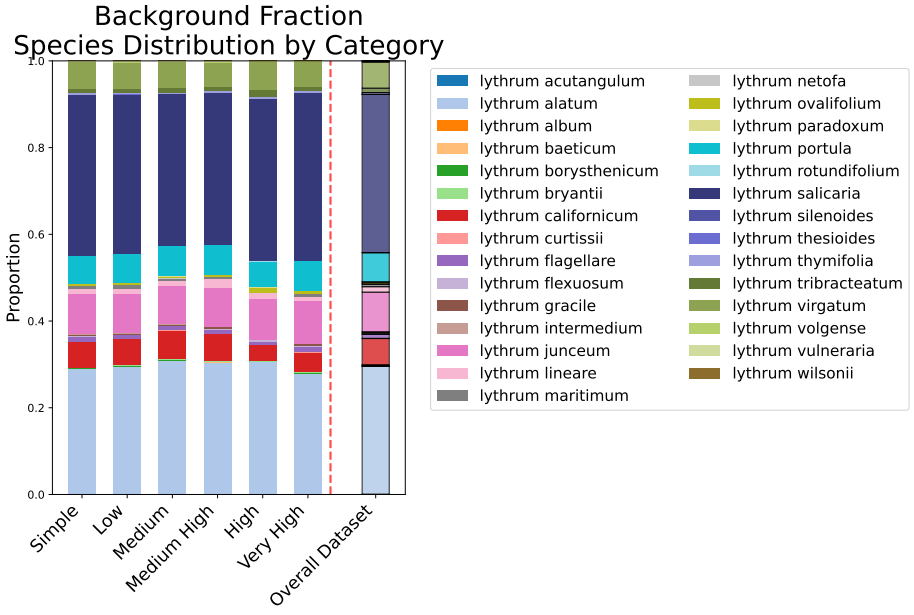
Table C.5: Results of Image complexity (i.e., the number of regions per image) correlation with accuracy analysis. For each region coverage category we report mean accuracy, sample size and KL divergence from the dataset distribution.

ANOVA: $F = 2.554$, $p = 0.026$; the highest mean accuracy is observed in the 'Very High' (10+) category (0.850) and the lowest in the 'Simple' (1) category (0.771). Accuracy differences are statistically significant ($p < 0.05$).

for the model, the observed effect is limited.

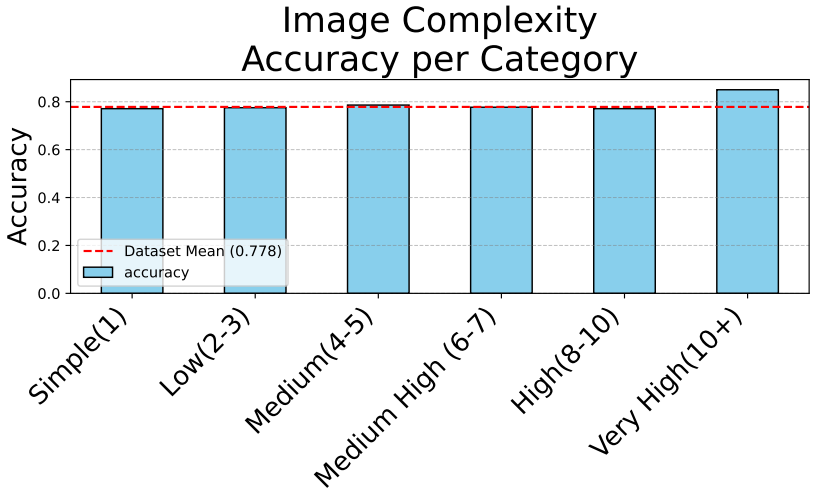


(a)

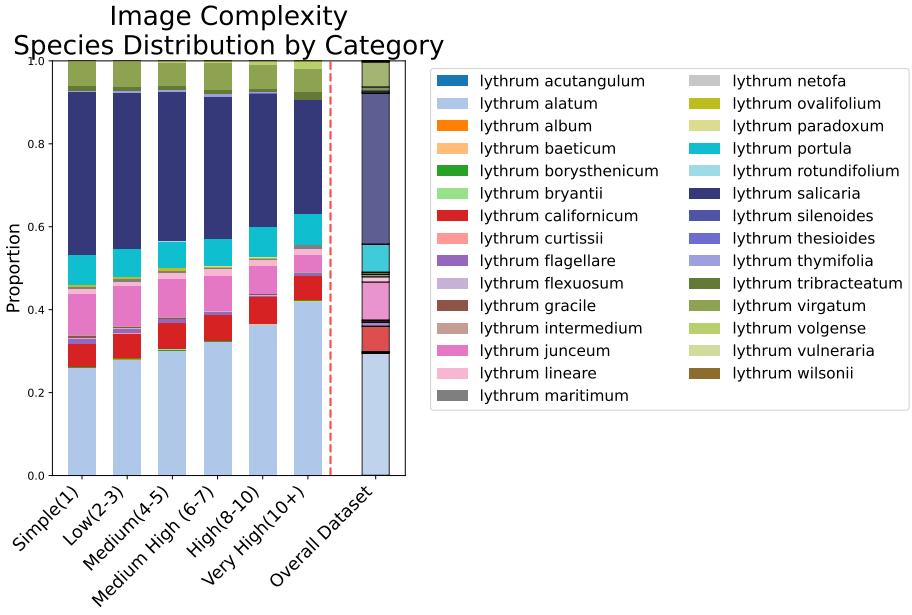


(b)

Figure C.4: For the Background/undefined fraction analysis we report accuracy for each category (a) and distribution of species for each category (b).



(a)



(b)

Figure C.5: For the Image complexity analysis we report accuracy for each category (a) and distribution of species for each category (b).

Bibliography

[1] M. van Kleunen, W. Dawson, and N. Maurel. «Characteristics of successful alien plants». In: *Molecular Ecology* 24.9 (2015), pp. 1954–1968. DOI: 10.1111/mec.13013 (cit. on pp. 4, 13).

[2] R. Mathakutha, C. Steyn, P. C. le Roux, I. J. Blom, S. L. Chown, B. H. Daru, B. S. Ripley, A. Louw, and M. Greve. «Invasive species differ in key functional traits from native and non-invasive alien plant species». In: *Journal of Vegetation Science* 30.5 (2019), pp. 994–1006. DOI: 10.1111/jvs.12772 (cit. on pp. 4, 5, 13).

[3] M. van Kleunen, E. Weber, and M. Fischer. «A meta-analysis of trait differences between invasive and non-invasive plant species». In: *Ecology Letters* 13.2 (2010), pp. 235–245. DOI: 10.1111/j.1461-0248.2009.01418.x (cit. on pp. 4, 13).

[4] Y. Li, M. Yue, Y. Wang, Z. Mao, J. Lyv, and Q. Li. «Invasive-plant traits, native-plant traits, and their divergences as invasion factors». In: *Ecology and Evolution* 14.6 (2024), e11525. DOI: 10.1002/ece3.11525 (cit. on p. 5).

[5] Alejandro Ordonez, Ian J Wright, and Han Olff. «Functional differences between native and alien species: a global-scale comparison». In: *Functional Ecology* 24.6 (2010), pp. 1353–1361 (cit. on p. 5).

[6] A. J. Leffler, J. J. James, T. A. Monaco, and R. L. Sheley. «A new perspective on trait differences between native and invasive exotic plants». In: *Ecology* 95.2 (2014), pp. 298–305. DOI: 10.1890/13-0102.1 (cit. on p. 5).

[7] W. Dawson, N. Maurel, and M. van Kleunen. «A new perspective on trait differences between native and invasive exotic plants: comment». In: *Ecology* 96.4 (2015), pp. 1150–1152. DOI: 10.1890/14-1315.1 (cit. on p. 5).

[8] Milan Šulc and Jiřı Matas. «Fine-grained recognition of plants from images». In: *Plant Methods* 13.1 (2017), p. 115 (cit. on p. 6).

[9] Voncarlos M Araújo, Alceu S Britto Jr, Luiz S Oliveira, and Alessandro L Koerich. «Two-view fine-grained classification of plant species». In: *Neurocomputing* 467 (2022), pp. 427–441 (cit. on p. 6).

[10] Matthew R Keaton, Ram J Zaveri, Meghana Kovur, Cole Henderson, Donald A Adjeroh, and Gianfranco Doretto. «Fine-grained visual classification of plant species in the wild: Object detection as a reinforced means of attention». In: *arXiv preprint arXiv:2106.02141* (2021) (cit. on p. 6).

[11] Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, and João V. B. Soares. «Leafsnap: A Computer Vision System for Automatic Plant Species Identification». In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 502–516. ISBN: 978-3-642-33709-3 (cit. on p. 6).

[12] Aydin Kaya, Ali Seydi Keceli, Cagatay Catal, Hamdi Yalin Yalic, Huseyin Temucin, and Bedir Tekinerdogan. «Analysis of transfer learning for deep neural network based plant classification models». In: *Computers and electronics in agriculture* 158 (2019), pp. 20–29 (cit. on p. 6).

[13] Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. «The iNaturalist Challenge 2017 Dataset». In: *CoRR* abs/1707.06642 (2017). arXiv: 1707.06642. URL: http://arxiv.org/abs/1707.06642 (cit. on p. 6).

[14] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. «The herbarium challenge 2019 dataset». In: *arXiv preprint arXiv:1906.05372* (2019) (cit. on p. 6).

[15] Riccardo de Lutio et al. «The herbarium 2021 half–earth challenge dataset and machine learning competition». In: *Frontiers in Plant Science* 12 (2022), p. 787127 (cit. on p. 7).

[16] Jose Carranza-Rojas, Hervé Goeau, Pierre Bonnet, Erick Mata-Montero, and Alexis Joly. «Going deeper in the automated identification of Herbarium specimens». In: *BMC evolutionary biology* 17.1 (2017), p. 181 (cit. on p. 7).

[17] Hervé Goëau, Pierre Bonnet, and Alexis Joly. «Overview of PlantCLEF 2022: Image-based plant identification at global scale». In: CEUR-WS. 2022 (cit. on p. 7).

[18] Hervé Goeau, Vincent Espitalier, Pierre Bonnet, and Alexis Joly. «Overview of PlantCLEF 2024: multi-species plant identification in vegetation plot images». In: CEUR-WS. 2024 (cit. on p. 7).

[19] Maxime Oquab et al. «Dinov2: Learning robust visual features without supervision». In: *arXiv preprint arXiv:2304.07193* (2023) (cit. on p. 7).

[20] Samuel Stevens et al. «Bioclip: A vision foundation model for the tree of life». In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 19412–19424 (cit. on pp. 7, 13–15, 21).

[21] Jianyang Gu et al. «Bioclip 2: Emergent properties from scaling hierarchical contrastive learning». In: *arXiv preprint arXiv:2505.23883* (2025) (cit. on pp. 7, 12, 14, 18, 21).

[22] Jackson Baron, DJ Hill, and H Elmiligi. «Combining image processing and machine learning to identify invasive plants in high-resolution images». In: *International Journal of Remote Sensing* 39.15-16 (2018), pp. 5099–5118 (cit. on p. 7).

[23] Tobias Jensen, Frederik Seerup Hass, Mohammad Seam Akbar, Philip Holm Petersen, and Jamal Jokar Arsanjani. «Employing machine learning for detection of invasive species using sentinel-2 and aviris data: The case of Kudzu in the United States». In: *Sustainability* 12.9 (2020), p. 3544 (cit. on p. 7).

[24] Thomas A Lake, Ryan D Briscoe Runquist, and David A Moeller. «Deep learning detects invasive plant species across complex landscapes using Worldview-2 and Planetscope satellite imagery». In: *Remote Sensing in Ecology and Conservation* 8.6 (2022), pp. 875–889 (cit. on p. 7).

[25] Reuben P Keller, Dragi Kocev, and Sašo Džeroski. «Trait-based risk assessment for invasive species: high performance across diverse taxonomic groups, geographic ranges and machine learning/statistical tools». In: *Diversity and Distributions* 17.3 (2011), pp. 451–461 (cit. on p. 8).

[26] Eva Grotkopp, Marcel Rejmánek, Michael J Sanderson, and Thomas L Rost. «Evolution of genome size in pines (Pinus) and its life-history correlates: supertree analyses». In: *Evolution* 58.8 (2004), pp. 1705–1729 (cit. on p. 8).

[27] Imageomics institute. *Imageomics*. [Online; accessed 8-September-2025]. 2025. URL: `https://imageomics.osu.edu/about` (cit. on p. 8).

[28] Moritz D Lürig, Seth Donoughe, Erik I Svensson, Arthur Porto, and Masahito Tsuboi. «Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology». In: *Frontiers in Ecology and Evolution* 9 (2021), p. 642774 (cit. on p. 8).

[29] Meghan A Balk et al. «A FAIR and modular image-based workflow for knowledge discovery in the emerging field of imageomics». In: *Methods in ecology and evolution* 15.6 (2024), pp. 1129–1145 (cit. on p. 8).

[30] Tanya Berger-Wolf. «HDR Institute: Imageomics: A New Frontier of Biological Information Powered by Knowledge-Guided Machine Learning». In: *NSF Award Number 2118240. Directorate for Computer and Information Science and Engineering* 21.2118240 (2021), p. 18240 (cit. on p. 8).

[31] NORMAN MACLEOD. «On the use of machine learning in morphometric analysis». In: *Biological shape analysis: proceedings of the 4th international symposium*. World Scientific. 2017, pp. 134–171 (cit. on p. 8).

[32] Moritz Lürig. *phenopype–a phenotyping pipeline for python (Version 0.4. 5)*. 2018 (cit. on p. 8).

[33] Arthur Porto and Kjetil L Voje. «ML-morph: A fast, accurate and general approach for automated detection and landmarking of biological structures in images». In: *Methods in Ecology and Evolution* 11.4 (2020), pp. 500–512 (cit. on p. 8).

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. «Learning deep features for discriminative localization». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929 (cit. on p. 9).

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. «Grad-cam: Visual explanations from deep networks via gradient-based localization». In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626 (cit. on p. 9).

[36] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. «Axiomatic attribution for deep networks». In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328 (cit. on pp. 9, 15).

[37] Vitali Petsiuk, Abir Das, and Kate Saenko. «Rise: Randomized input sampling for explanation of black-box models». In: *arXiv preprint arXiv:1806.07421* (2018) (cit. on p. 9).

[38] Scott M Lundberg and Su-In Lee. «A unified approach to interpreting model predictions». In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 9, 15).

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. «" Why should i trust you?" Explaining the predictions of any classifier». In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144 (cit. on p. 9).

[40] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. «Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)». In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677 (cit. on p. 9).

[41] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. «This looks like that: deep learning for interpretable image recognition». In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 9).

[42] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. «Deformable protopnet: An interpretable image classifier using deformable prototypes». In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10265–10275 (cit. on p. 9).

[43] Masato Shirai et al. «Development of a system for the automated identification of herbarium specimens with high accuracy». In: *Scientific Reports* 12.1 (2022), p. 8066 (cit. on p. 9).

[44] Jihen Amara, Birgitta König-Ries, and Sheeba Samuel. «Explainability of Deep Learning-Based Plant Disease Classifiers Through Automated Concept Identification». In: *arXiv preprint arXiv:2412.07408* (2024) (cit. on p. 9).

[45] Cody E Hinchliff et al. «Synthesis of phylogeny and taxonomy into a comprehensive tree of life». In: *Proceedings of the National Academy of Sciences* 112.41 (2015), pp. 12764–12769 (cit. on p. 13).

[46] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. «Benchmarking representation learning for natural world image collections». In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12884–12893 (cit. on p. 14).

[47] Zahra Gharaee et al. «A step towards worldwide biodiversity assessment: The BIOSCAN-1M insect dataset». In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 43593–43619 (cit. on p. 14).

[48] Alec Radford et al. «Learning transferable visual models from natural language supervision». In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763 (cit. on p. 14).

[49] Leland McInnes, John Healy, and James Melville. «Umap: Uniform manifold approximation and projection for dimension reduction». In: *arXiv preprint arXiv:1802.03426* (2018) (cit. on p. 18).

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep residual learning for image recognition». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 21).

[51] S. Lowe, M. Browne, S. Boudjelas, and M. De Poorter. *100 of the World's Worst Invasive Alien Species: A Selection from the Global Invasive Species Database*. Published by the Invasive Species Specialist Group (ISSG), a specialist group of the Species Survival Commission (SSC) of the IUCN. First published as special lift-out in Aliens 12, December 2000. Updated and reprinted version: November 2004. Auckland, New Zealand, 2000, p. 12 (cit. on p. 29).

[52] Royal Botanics Garden. *Plants of the World Online*. [Online; accessed 9-September-2025]. 2025. URL: https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:30002863-2 (cit. on p. 29).

[53] Bernd Blossey, Luke C Skinner, and Janith Taylor. «Impact and management of purple loosestrife (Lythrum salicaria) in North America». In: *Biodiversity & Conservation* 10.10 (2001), pp. 1787–1807 (cit. on p. 29).

[54] Global Invasive Species Database. *Species profile: Lythrum salicaria*. http://www.iucngisd.org/gisd/speciesname/Lythrum+salicaria. Downloaded on 23-09-2025. 2025 (cit. on p. 29).

[55] New Zealand Plant Conservation Network. *Lythrum hyssopifolia*. https://www.nzpcn.org.nz/flora/species/lythrum-hyssopifolia/. Accessed on: 23-09-2025. 2025 (cit. on p. 29).

[56] Invasives Foundation (South Africa). *Hyssop Loosestrife Fact Sheet*. https://invasives.org.za/fact-sheet/hyssop-loosestrife. Accessed on: 23-09-2025. 2025 (cit. on p. 29).

[57] Kali Z Mattingly, Brenna N Braasch, and Stephen M Hovick. «Greater flowering and response to flooding in Lythrum virgatum than L. salicaria (purple loosestrife)». In: *AoB Plants* 15.2 (2023), plad009 (cit. on p. 29).

[58] Fraser Valley Invasive Species Society. *Wand Loosestrife (*Lythrum virgatum*)*. https://fviss.ca/invasive-plant/wand-loosestrife. Accessed on: 23-09-2025. 2025 (cit. on p. 29).

[59] iNaturalist community. *Observations of ['lythrum acutangulum', 'lythrum alatum', 'lythrum album', 'lythrum baeticum', 'lythrum borysthenicum', 'lythrum bryantii', 'lythrum californicum', 'lythrum curtissii', 'lythrum flagellare', 'lythrum flexuosum', 'lythrum gracile', 'lythrum hyssopifolia', 'lythrum intermedium', 'lythrum junceum', 'lythrum lineare', 'lythrum maritimum', 'lythrum netofa', 'lythrum ovalifolium', 'lythrum paradoxum', 'lythrum portula', 'lythrum rotundifolium', 'lythrum salicaria', 'lythrum silenoides', 'lythrum thesioides', 'lythrum thymifolia', 'lythrum tribracteatum', 'lythrum virgatum', 'lythrum volgense', 'lythrum vulneraria', 'lythrum wilsonii'] from [Afghanistan, Alabama, Albania, Algeria, Altay, Amur, Argentina Northeast, Argentina Northwest, Arizona, Arkansas, Austria, Azores, Baleares, Baltic States, Belarus, Belgium,*

Bolivia, Brazil South, Bulgaria, Buryatiya, California, Canary Is., Central European Russia, Chad, Chile Central, Chile North, China North-Central, China South-Central, Chita, Colombia, Colorado, Connecticut, Corse, Cuba, Cyprus, Czechia-Slovakia, Delaware, Denmark, District of Columbia, Dominican Republic, DR Congo, East Aegean Is., East European Russia, Ecuador, Egypt, Ethiopia, Finland, Florida, France, Georgia, Germany, Great Britain, Greece, Guatemala, Haiti, Hawaii, Hungary, Illinois, Indiana, Inner Mongolia, Iowa, Iran, Iraq, Ireland, Irkutsk, Italy, Japan, Juan Fernández Is., Kansas, Kazakhstan, Kentucky, Kenya, Khabarovsk, Kirgizstan, Korea, Krasnoyarsk, Kriti, Krym, Kuril Is., Lebanon-Syria, Libya, Louisiana, Madeira, Maine, Malawi, Manchuria, Maryland, Massachusetts, Mexico Central, Mexico Gulf, Mexico Northeast, Mexico Northwest, Mexico Southeast, Mexico Southwest, Michigan, Minnesota, Mississippi, Missouri, Mongolia, Morocco, Nebraska, Netherlands, Nevada, New Hampshire, New Jersey, New Mexico, New South Wales, New York, North Carolina, North Caucasus, North Dakota, North European Russia, Northern Territory, Northwest European Russia, Norway, NW. Balkan Pen., Ohio, Oklahoma, Ontario, Pakistan, Palestine, Pennsylvania, Peru, Poland, Portugal, Primorye, Qinghai, Queensland, Rhode I., Romania, Rwanda, Sakhalin, Sardegna, Saudi Arabia, Senegal, Sicilia, Sinai, Socotra, Somalia, South Australia, South Carolina, South Dakota, South European Russia, Spain, Sudan-South Sudan, Sweden, Switzerland, Tadzhikistan, Tanzania, Tasmania, Tennessee, Texas, Tibet, Transcaucasus, Tunisia, Turkmenistan, Tuva, Türkey, Türkey-in-Europe, Uganda, Ukraine, Uruguay, Utah, Uzbekistan, Venezuela, Vermont, Victoria, Virginia, West Himalaya, West Siberia, West Virginia, Western Australia, Wisconsin, Wyoming, Xinjiang, Yemen, Alaska, Alberta, Argentina South, British Columbia, Cape Provinces, Chile South, Idaho, Manitoba, Montana, New Brunswick, New Zealand North, Newfoundland, Norfolk Is., Nova Scotia, Oregon, Prince Edward I., Québec, Saskatchewan, Washington] observed between [21/07/1940 - 19/06/2025]. `https://www.inaturalist.org`. Exported from iNaturalist on [20/06/2025]. 2025 (cit. on p. 29).

[60] WFO. *Lythrum salicaria subsp. intermedium (Fisch. ex Colla) H. Hara.* `http://www.worldfloraonline.org/taxon/wfo-0001076019`. Published on the Internet. Accessed on: 23 Sep 2025. 2025 (cit. on p. 33).