Politecnico di Torino

Politecnico di Torino

Data Science and Engineering

A.Y. 2024/2025

Graduation Session October 2025

Explainability-Driven Deep Learning for Predicting Biological Invasiveness in Plant Species

Relatori:

Daniele Apiletti

Simone Monaco

Candidati:

Barbara Frittella

Abstract

Invasive species spread rapidly outside their natural range and can disrupt ecosystems, damage economies, and threaten health. Early identification is therefore critical, yet current ecological practice remains largely manual and existing deep learning pipelines provide little insight into the morphological traits that define invasiveness. This thesis addresses this gap by developing an explainability-driven deep learning pipeline that links morphological traits to a model's classification prediction of invasiveness and exposes why the model fails on specific images. The approach trains a classifier on image embeddings extracted with BioCLIP-2 and adopts an imageomics perspective, treating images as high-dimensional phenotypes. Saliency-guided region extraction (Integrated Gradients) identifies the image pixels most critical to the model's predictions. By clustering the embeddings of these regions and manually annotating them, we are able to define interpretable visual concepts. These clusters are then independently validated and propagated across the dataset to produce image-level concept labels. By analyzing the co-occurrence between labels and model outputs, the pipeline identifies which structures support correct invasive detections and which spurious cues drive misclassifications. This method is demonstrated on a dataset of images of *Lythrum*, a plant genus of around 40 species in the family Lythraceae, offering a scalable path toward more transparent and trustworthy deep learning systems in ecology. Future works could extend the pipeline beyond *Lythrum* to other taxa of ecological importance, also integrating complementary data sources to link image-derived concepts more directly to measurable biological traits.

Acknowledgements

I would like to express my deepest gratitude to Dr.Simone Monaco for his guidance, patience, and continuous support throughout this work. His advice influenced every stage of the research, from the experimental design to the final writing of this thesis.

I'm also thankful to Dr.Riccardo Ciarle for his help and for sharing his expertise on the biological aspects of this research. His contributions were crucial in keeping our analysis consisten and in avoiding mistakes when addressing complex biological issues.

Lastly, I would like to thank my colleague and research partner, Guido Spina, with whom I shared every step of this project. Working side by side made this experience not only more productive but also lighter and much more enjoyable. The long days seemed shorter, and the challenges easier to face.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Invasive species are organisms that humans have transported, intentionally or not, outside their natural range. Once they settled in a new environment, they spread quickly and disrupt existing ecosystems.

However, not every species that is moved to a new environment becomes invasive. To considered invasive, a species must first be classified as 'alien' (that is, non-native to a specific region), and only a portion of these alien species turn out to be invasive.

Preventing the arrival and spread of invasive species is vital and limiting their effects is crucial to avoid the deterioration of ecosystems. Failing to do so can lead to significant problems for the economy, food security, and even human health.

A well-known example of disruption caused by an invasive species is that of the the water hyacinth (*Pontederia crassipes*). Native to South America, this floating aquatic plant serves as a key food source for the Amazonian manatee, which naturally limits its growth. When introduced into other ecosystems, however, the water hyacinth has a strong impact: it can outcompete native aquatic plants, hindering their photosynthesis and growth. By covering the surface of the water, it blocks sunlight from reaching underwater vegetation, leading to their death. The decomposition of this plant material then consumes large amounts of oxygen, causing oxygen depletion and eventually resulting in fish kills.

The excessive presence of water hyacinth can also cause several health issues for humans. This plant is capable of absorbing considerable amounts of heavy metals and other substances that are toxic to people. When the plant dies and decays, it releases these compounds back into the environment, polluting the water and reducing its quality, sometimes even contaminating drinking water for nearby populations.

The economic consequences of invasions by *Pontederia crassipes* are also highly significant. One of the main reasons is that dense infestations in water bodies (such as rivers or lakes) can severely limit or completely block transportation, both for people and goods. In the United States, it has earned the nickname *'million dollar*

weed', not because of its worth, but due to the enormous sums of money spent every year by local authorities on its removal, which is often only partially successful.

For these reasons, it is essential to develop the ability to accurately identify invasive species, either to reduce the consequences of an ongoing invasion or to prevent one from occurring in the first place.

Many ecological studies have explored what makes certain alien species invasive. These works focus on identifying functional traits (such as growth rate, dispersal strategy, or seed mass) that allow a species to dominate others once introduced into new environments. While this approach can predict invasiveness, it faces key challenges: many studies cover only small groups of species or specific regions, and they depend on biological or ecological data that can be hard to obtain consistently across taxa.

To our knowledge, no previous work has focused on predicting invasiveness using only image data. This is the central aim of our study: to design a pipeline capable of estimating the potential invasiveness of a species within the same genus, based solely on visual traits.

As a case study, we selected the genus *Lythrum*, part of the *Lythraceae* family, which includes annual and perennial herbaceous species. This genus was chosen because one of its members (*Lythrum salicaria*, commonly known as purple loosestrife) is listed by the IUCN (International Union for Conservation of Nature) among the '100 of the World's Worst Invasive Alien Species'. Native to Europe, Asia, northern Africa, and eastern Australia, *L. salicaria* has invaded wetland ecosystems in North America, where it outcompetes native flora.

We collected images of *Lythrum* species from *iNaturalist*, a citizen science platform where users upload and label photographs of living organisms. The genus is well represented on this platform: we gathered images of 30 *Lythrum* species, three of which are recognized as invasive outside their native range (*Lythrum hyssopifolia, Lythrum salicaria* and *Lythrum virgatum*).

One of the main tools used in this work is BioCLIP, a state-of-the-art vision model trained on biological data. It is based on OpenAI's CLIP framework and trains a vision encoder and a text encoder together through contrastive learning using image-text pairs. BioCLIP has demonstrated the ability to extract detailed representations from image data, capturing subtle biological structures and distinguishing between species that have a similar appearance or are poorly represented in the training set.

In our study, we employ BioCLIP as a feature extractor to map images from the dataset into multi-dimensional embeddings. These embeddings are then used to train a classifier that can differentiate invasive species from non-invasive ones.

To improve the interpretability of the workflow, we introduce an explainability component. For each image, we generate attribution maps using Integrated Gradients, an algorithm designed to highlight the regions that most influence the model's predictions. We then extract these relevant regions and group them into

clusters to determine which morphological structures they correspond to. This approach allows us to identify the distinctive traits of each species that appear in the images and to connect them with the model's predictions.

Finally, we analyze the resulting data to uncover recurring patterns and to better understand which visual morphological features affect a species' likelihood of being invasive or non-invasive.

This thesis is organized as follows:

- **Chapter 2** reviews the existing literature, covering ecological studies on invasive traits, applications of deep learning to plant identification, and recent work on explainability models.
- **Chapter 3** describes the full methodology, including the classification model, the explainability pipeline, and the clustering analysis.
- **Chapter 4** presents experimental results on the *Lythrum* dataset, discussing classification accuracy, saliency maps, and the morphological patterns discovered.
- **Chapter 5** concludes the work, summarizing contributions, acknowledging limitations, and outlining future directions such as applying the method to other taxa or integrating complementary trait data.

The two main contributions of this thesis are:

- We demonstrate that invasiveness can be inferred from visual traits alone, without relying on tabular or categorical ecological data;
- We introduce an interpretable framework that connects computer vision and ecology, offering a foundation for future research that combines automated image analysis with traditional biological knowledge.

Chapter 2

Related Works

2.1 Biological approach to the identification of traits related to invasiveness

The study of traits that help invasive plant species survive beyond their native ranges has long attracted in ecology. Researchers aim to identify which traits predict whether a plant will become invasive.

Van Kleunen et al. (2015) [1] propose a framework consisting of a set of questions aimed at understanding the success of alien species. Each question depends on the answer to the previous one, reflecting the hierarchical structure of the invasion process. The sequence of questions progresses from broader geographical regions to smaller communities, and each step includes a group of traits that are considered to influence the success of alien species.

Mathakutha et al. (2019) [2] focused on functional traits and asked two main questions. Are invasive plants functionally different from native species? Which traits set invasives apart from non-invasive aliens? They found that most traits differed between invasive and native plants, suggesting that functional traits relate to invasion success. The traits most tied to invasiveness were plant height, leaf area, frost tolerance, and specific leaf area. Their dataset was small though (13 traits measured across 26 species from 13 families in the sub-Antarctic region), so results should be read with caution.

A broader view comes from Van Kleunen et al. (2010) [3], who ran a meta-analysis to test links between invasiveness and performance traits such as growth rate, leaf allocation, and fitness. They compared 125 invasive with 196 non-invasive species. Across all traits, invasive species differed more from native species than from other alien species. When compared with native species that are invasive elsewhere, no trait differences were significant. The authors concluded that future invasions might be predicted from measurable species traits.

Li et al. (2024) [4] examined how the traits of both invasive and native plants interact during invasion. Their results show that the outcome depends on the species pairing. For instance, native plants such as *Artemisia argyi, Artemisia lavandulifolia* and *Chenopodium album* were more competitive when paired with certain invasives. In contrast, *Setaria viridis, Austrocylindropuntia vestita*, and *Artemisia annua* outcompeted *Elodea canadensis, Galinsoga quadriradiata*, and *Erigeron annuus* respectively. They found that invasive success was linked mainly to plant height, stem diameter, and biomass allocation. Native competitiveness depended on biomass distribution, stem size, and functional group differences.

Both Mathakutha et al. [2] and Li et al. [4] tested ideas first outlined by Ordonez et al. (2010) [5]: the 'phenotypic divergence' and 'phenotypic convergence' hypotheses. The divergence view holds that invaders succeed because they differ from natives, exploiting open ecological niches. The convergence view suggests that shared environmental pressures force similar traits in both groups. Evidence from both studies ([2, 4]) supports phenotypic divergence. Distinct functional traits, rather than shared ones, appear to promote invasion.

Leffler et al. (2014) [6] offered a different take. They argued that the meaning of trait differences depends on ecological context and warned that predicting invasion from traits alone would be difficult. They proposed a rule: differences between a native and an exotic invasive species must exceed the range of variation found among co-occurring natives to be meaningful. Dawson et al. (2015) [7] later challenged this view. They pointed out that Leffler's rule cannot distinguish between cases where alien traits fall within native variation but still function differently and cases where traits are nearly identical.

These and other studies offer valuable insights for understanding the differences between native species, non-invasive alien species, and alien species that become successfully invasive. They also help in predicting whether a species is likely to become invasive when introduced into a new environment. However, although several works have shown that the invasion success of plant species can be inferred from their functional traits, many of them face important challenges and limitations. Among these are experimental conditions involving only a small number of plants or studies confined to specific regions, which makes it difficult to extend the findings to larger scales or to datasets with greater complexity. This raises the question of whether such analyses could be made scalable through the use of deep learning methods.

2.2 Deep learning for plant species identification

The development of systems for automatically recognizing plant species from images is a critical interdisciplinary problem that bridges computer vision and biodiversity

science. This challenge falls under fine-grained visual classification, where models must distinguish among thousands of species based on subtle morphological features, such as leaf shape, venation, and flower structure. This complexity is further intensified by large variability within species (induced by growth stage and environmental conditions) and by strong visual similarity between species, especially among closely related taxa [8, 9, 10].

Early approaches to automated plant species identification relied on predefined feature engineering. Systems such as *Leafsnap* [11] employed traditional computer vision techniques, extracting histograms of curvatures along the contour of the leaf, to match against curated references. These methods successfully demonstrated the potential of automated recognition, but they were limited in scalability and struggled with noisy, real world data. The introduction of convolutional neural networks (CNNs) revolutionized the field: by learning features directly from raw pixels, CNNs rapidly outperformed traditional methods, establishing deep learning as the dominant approach for fine-grained plant classification. Given the limited size of many botanical datasets, researchers quickly adopted transfer learning from pretrained models [12], which provided strong generalization even with limited training examples.

This progress has been critically enabled by the release of large, real world datasets, of which the iNaturalist collection is a primary example. Presented by Van Horn et al.[13], this dataset contains over 859000 images across more than 5000 species generated by a global community of citizen scientists, capturing the visual diversity and ecological realism inherent in real-world observations. Baseline performance on this dataset only achieved 67% top-1 accuracy, with severe degradation on rare classes, highlighting the challenges of long-tail distributions.

Herbarium specimens serve as a complementary data source, providing a standardized yet fine-grained and taxonomically rich alternative. The potential of deep learning in this domain was demonstrated by the Herbarium 2019 Challenge [14], a dataset of more than 46000 expertly labeled images of the Melastomataceae family across 683 species. Top performers in the associated FGVC6 competition achieved 89.8% classification accuracy. The Herbarium 2021 dataset [15] expanded to over 2.5 million images of 64500 taxa, particularly challenging for its pronounced class imbalance (imbalance factor > 1650) and its broad representation of major plant divisions. In the associated competition, model performance was measured using the F1 score, and the best submitted result was 0.757. Carranza-Rojas et al. [16] explored the application of CNNs to herbarium sheets. This analysis provided valuable insights into the effective use of transfer learning in this domain, identifying when its application is effective or counterproductive.

The PlantCLEF series has played a key role in driving progress in plant recognition by posing increasingly complex challenges. The 2022 edition [17], for example, focused on identifying 80000 species from 4 million images from heterogeneous

sources. The documentation mentions the difficulty of building models that can generalize across different data quality and types. The 2024 competition [18] introduced the task of identifying multiple species within single vegetation plot images and treated plant identification as a weakly-labeled multi-label classification task. Two pre-trained models were shared to solve this problem, both based on Vision Transformer (ViT) architecture initially pretrained with the DinoV2 Self-Supervised Learning approach [19].

More recently, BioCLIP [20] introduced a foundation model trained on a dataset with 10 million biology images, using hierarchical contrastive learning with taxonomic labels, and establishing a new state-of-the-art in the fine-grained biology classification task. Its successor, BioCLIP 2 [21], scaled this approach using 214 million images.

2.3 Deep learning for identification of traits related to invasiveness

Many researchers have used deep learning to detect invasive species and reduce the harm they cause to ecosystems.

Baron et al. (2018) [22] combined image processing with machine learning to identify yellow flag iris (*Iris pseudacorus*, an invasive species) from drone images. Jensen et al. (2020) [23] applied several machine learning classifiers to map kudzu vine (*Pueraria montana*, an invasive species) across the southeastern United States using spatial data. Likewise, Lake et al. (2022) [24] used WorldView-2 and PlanetScope satellite imagery with convolutional neural networks to detect leafy spurge (*Euphorbia virgata*, an invasive species) across complex landscapes. These studies focused on locating known invasive species in the environment, not on identifying the biological traits that make them invasive.

Focusing on the identification of traits associated with invasive species, Keller et al. (2011) [25] examined trait-based risk assessments for invasive species by using six different datasets that range from regional to global scales and cover various taxa, regions, and invasion stages. Among these six datasets, two refer to birds, two to fish, one to molluscs, and one to pines. For the latter (Pinus [26]), the authors considered several categorical and numerical traits, identifying seed mass, dispersal mode, serotiny, generation time, reproductive intervals, fire tolerance, and environmental tolerances as the main predictive traits. They also compared two statistical approaches and seven machine learning algorithms, finding no significant difference in the results produced by the two types of methods.

More recently, with advances in deep learning, the concept of *imageomics* has emerged [27]. This field aims to extract biological traits directly from images by embedding structured biological knowledge into learning algorithms. Within it,

phenomics focuses on identifying phenotypic traits from image data [28, 29, 30].

In this context, Macleod (2017) [31] compared traditional geometric morphometric techniques with more recent machine learning approaches for the analysis of digital images of carnivore skulls. The study assessed how effectively each method was able to describe group differences and examined their suitability for morphometric analysis.

Lürig et al. (2018) [32] designed a pipeline aimed at simplifying the immediate extraction of high-dimensional phenotypic data from digital images, enabling biologists to concentrate on fast and reproducible data collection.

Similarly, Porto et al. (2020) [33] proposed a machine learning pipeline for extracting detailed morphometric data from two-dimensional images of semi-rigid biological structures.

Previous studies have largely concentrated on numerical or categorical tabular traits (for example, seed mass or dispersal mode), using statistical or machine learning techniques for their analysis. However, this type of approach may overlook morphological characteristics that can only be observed visually, such as petal color or stem structure.

At the same time, work using deep learning has focused mainly on species identification, not on trait detection. To our knowledge, no published research has yet used deep learning to identify visual morphological traits linked to the potential invasiveness of plants.

2.4 Explainability mentions

The “black box" nature of deep learning models poses a challenge for their adoption in ecological domains, where both accuracy and trustworthiness are critical. To be reliable, predictions must be explainable, based on scientifically valid features rather than spurious correlations. In fine-grained classification tasks, this means verifying that models focus on biologically meaningful traits, such as leaf shape and flower morphology, rather than irrelevant cues like lighting or background noise. For critical applications like invasive species detection, explainability ensures that models use true diagnostic traits, preventing misinformed decisions.

Explainability methods in computer vision generally fall into a few established families:

- **Saliency and gradient-based methods**. Methods like CAM [34] and Grad-CAM [35] produce heatmaps that highlight image regions responsible for a given prediction by using class-specific gradients over the final convolutional layer. Integrated Gradients [36] instead traces a linear path from a baseline (e.g., a blank image) to the actual input and integrates the gradients along that path to assess feature relevance.

- **Perturbation-based methods**. These methods analyze a model's behavior by modifying its input. RISE [37] empirically identifies important pixels by applying randomized masks and aggregating the corresponding model outputs. SHAP [38] assigns a theoretical importance score to each input feature by leveraging Shapley values from cooperative game theory. LIME [39] explains individual predictions by generating perturbed samples around the input and fitting a simple interpretable model (e.g., linear regression) to approximate the local behavior of the complex model.

- **Concept-based and prototype methods.** This family is designed to offer human-comprehensible reasoning, moving beyond simple feature attribution. TCAV [40] quantifies the influence of user-defined concepts on model decisions. ProtoPNet [41], and its successors, Deformable ProtoPNet[42], learn prototypical parts during training; predictions rely on similarity to these prototypes, offering example-based explanations.

These explainability techniques have already shown promise in ecological applications. For instance, a study on herbarium specimen classification used Grad-CAM to reveal that models often mimic expert behavior: first observing the overall plant, then focusing on diagnostic parts, confirming that the model indeed pays attention to biologically meaningful regions [43]. A more recent work applied concept-based methods to plant disease detection, revealing both meaningful visual cues and spurious biases [44]. Despite these promising examples, the systematic application of explainability in biodiversity (and invasive species detection and monitoring) is still underexplored. Moreover, existing methods face challenges such as instability, the risk of highlighting visually salient but biologically irrelevant features, and the absence of quantitative standards for validation. These problems are amplified in ecological datasets due to complex backgrounds and metadata leakage. In this research, explainability serves as a critical validation tool, ensuring that models rely on biologically meaningful traits and that their predictions can be trusted for scientific purposes.

2.5 Research question

In conclusion, while multiple research projects have investigated the identification of morphological traits linked to the invasiveness potential of plants, we are not aware of any study that relied solely on visual traits obtained from image data. By concentrating on the plant genus *Lythrum* (Lythraceae) and applying deep learning methods in a systematic manner, our goal is to distinguish invasive from non-invasive species based on their specific morphological characteristics.

Chapter 3

Methods

We designed a pipeline to extract information on morphological traits from plant images. It comprises three main modules, each containing several intermediate steps:

1. **Classification model** to predict whether an image represents an invasive or non-invasive species.
    - Extraction of **image embeddings** using a foundation model fine-tuned for the biological domain (BioCLIP 2);
    - **Training the classifier** on the extracted embeddings;
2. **Explainability pipeline:**
    - **Generation of heatmaps** to highlight the regions that most influence the model predictions using Integrated Gradients as the XAI algorithm;
    - **Extraction of regions** from the heatmaps by selecting bounding boxes that encompass the most influential pixels;
    - **Clustering phase:**
        - **Clustering of the regions** after embedding each region individually and reducing them to a two-dimensional space with UMAP;
        - **Cluster annotation** to identify the biological structure represented by each region (*Leaf, Flower, Stem*);
3. **Final analysis:** we analyze patterns and co-occurrences to discover the visual traits that drive a species' predicted invasiveness.

In the following sections, we will describe each step of the pipeline in detail, explaining the methodology and the decisions made throughout the research process.

3.1 Classification model

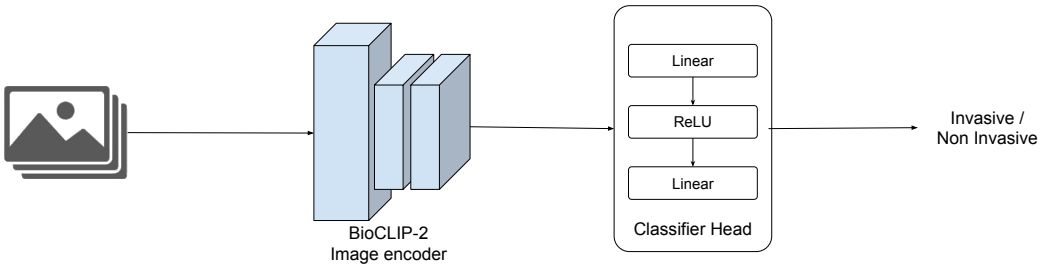


Figure 3.1: Classification model architecture.

In this section, we describe the procedure used to build the classification model. The model takes an image from the dataset (Sec. 4.1) as input and predicts whether it represents an *invasive* or *non-invasive* species. The complete architecture of the classification pipeline is shown in Fig. 3.1.

First, we used BioCLIP 2 [21] as an embedding extractor, employing the model's image encoder to generate a multidimensional embedding for each image in the dataset. These embeddings were then used to train a classifier. The image encoder was not fine-tuned on our dataset, so we assessed its ability to produce meaningful representations directly. While other embedding extractors could have been used, BioCLIP is specifically trained for the biological domain and provides better performance than a generic extractor, removing the need for additional training or fine-tuning.

The classifier, which takes the embedding as input, consists of a Linear layer, a ReLU activation, and a final Linear layer with two outputs corresponding to invasive and non-invasive classes.

We trained the classifier to predict whether an image, represented by its embedding, corresponds to an *invasive* or *non-invasive* species, using cross-entropy as the loss function. To handle the slight class imbalance in the dataset, class weights were incorporated into the cross-entropy loss. These weights were calculated as the inverse of the logarithm of the class sample counts, providing a balance between compensating for underrepresented classes and avoiding excessive weighting of rare categories. The cross-entropy loss formula and the method used to compute class weights are presented in Eq. 3.1.

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

Further details on the embedding models considered, the evaluation metrics, and the classifier's hyperparameters are provided in Sec. 3.4.1.

The classification task presents an inherent challenge. Invasiveness is not an absolute property: some species may behave invasively in one region and not in another. Researchers often differ in how they frame this distinction. Some studies compare invasive alien species with native species, aiming to isolate the traits that allow a non-native species to succeed locally. Other research contrasts only non-native species, comparing successful invaders against non-invasive aliens to identify the features that enable widespread establishment [1, 3, 2].

This study does not investigate the correlation between species traits and geographical distribution, but rather focuses on the relationship between a species' appearance and its invasiveness potential. Accordingly, we take a broader approach by considering any species that is invasive in at least one location as 'invasive' (at least potentially), and any species that is not invasive anywhere as 'non-invasive'. For this reason, we do not make distinctions between 'native' and 'non-native' species, concentrating exclusively on the potential for a species to become invasive.

3.1.1 BioCLIP

This section outlines what BioCLIP is and explains why we selected it as the feature extraction model for our study.

BioCLIP is a domain-specific vision-language foundation model designed to generalize across the entire tree of life [45]. It captures visual features that span diverse taxa and can distinguish among organisms that are morphologically similar. Another advantage is its robustness in low-data settings: BioCLIP can generate informative embeddings even for species that are underrepresented (or entirely absent) in the training data.

The model was trained on TreeOfLife-10M [20], a large, diverse dataset built by integrating multiple existing biological image collections, including iNat21 [46], BIOSCAN-1M [47] and the Encyclopedia of Life[1].

In terms of architecture, BioCLIP is based on OpenAI's CLIP framework [48], which uses transformer architectures and relies on the *self-attention* mechanism to capture contextual relationships between elements in a sequence. In the vision domain, as in this case, this concept is implemented through Vision Transformers (ViT), which divide an image into patches (instead of tokens, as in a standard transformer), serialize these patches into vectors, and process them through a transformer encoder like regular tokens. In BioCLIP, the vision encoder is a ViT-B/16, while the text encoder is a 77-token causal autoregressive transformer. Both encoders map their inputs into a shared embedding space, allowing for the measurement of similarity.

[1] https://eol.org

CLIP (and consequently BioCLIP) relies on a contrastive learning objective. During training, the model receives batches of paired image-text samples and learns to maximize similarity between matching pairs while minimizing it for mismatched ones. The contrastive loss function is expressed 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 a batch, and τ is a learnable temperature parameter [48].

In BioCLIP, the text encoder takes as input various combinations of common names, scientific names, and taxonomic names (Tab. 3.1 shows examples of these combinations as described in the original paper by Stevens et al. [20]). This multi-level textual supervision enhances the contrastive alignment process: the model does not simply learn to match an image with a single label, but also captures the semantic relationships across different linguistic representations, providing greater flexibility during testing.

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].

We decided to use BioCLIP in our study because it is pre-trained, avoiding the need for a long and computationally intensive training process, on a large dataset that extensively covers the species we are considering. We also take advantage of its ability to generate fine-grained image representations (important in our case, as several species have very similar morphologies) and its capacity to produce meaningful embeddings even for species that are rarely represented in the dataset (as is the case here, see Fig. 4.1).

In this work, we use the most recent version, BioCLIP-2 [21]. It incorporates a stronger vision transformer and is trained on TreeOfLife-200M, an expanded dataset built with the same structure as TreeOfLife-10M but on a much larger scale.

3.2 Explainability pipeline

We developed an explainability pipeline to provide insight into our classification model's decision process. Our goal is to validate whether the model's prediction of plant invasiveness is based on meaningful biological traits (and which ones specifically) or is influenced by spurious correlations like background noise. To achieve this, we designed the following procedure:

1. Generate heatmaps of model predictions to identify the image regions that drive classification decisions;

2. Extract the highlighted regions from the original images based on the heatmaps;

3. Label each extracted region to determine the biological structures it represents.

This process is shown in Fig. 3.2.

3.2.1 Heatmap generation

To ensure a comprehensive evaluation of feature attributions, we applied two complementary techniques for generating heatmaps: **Integrated Gradients** [36] and **Gradient SHAP**, a variant of SHAP [38]. These methods were chosen to cover different approaches to explainability: Integrated Gradients provides a deterministic, path-based gradient attribution, while Gradient SHAP introduces randomness and connects to Shapley values for fair feature attribution.

Integrated Gradients is a saliency and gradient-based method that computes feature attributions by integrating the gradients of the model's output with respect to the input along a straight path from a baseline to the input. This approach satisfies key axioms such as Sensitivity and Implementation Invariance, while requiring only a modest number of gradient evaluations for implementation. Intuitively, it measures how much each pixel contributes to moving the prediction away from the baseline.

Gradient SHAP builds on the same principle but introduces randomness: it samples multiple noisy versions of a baseline input and computes expectations of integrated gradients with respect to these perturbations. This formulation is theoretically linked to Shapley values from cooperative game theory, which guarantee a fair attribution of the model output among features.

Both methods were applied directly to our full classification pipeline (i.e., frozen BioCLIP-2 image encoder followed by the classifier head). This was *Step 1* of Fig. 3.2. Following a comparative analysis of their clustering outcomes, one method was chosen for final heatmap generation based on clustering performance.

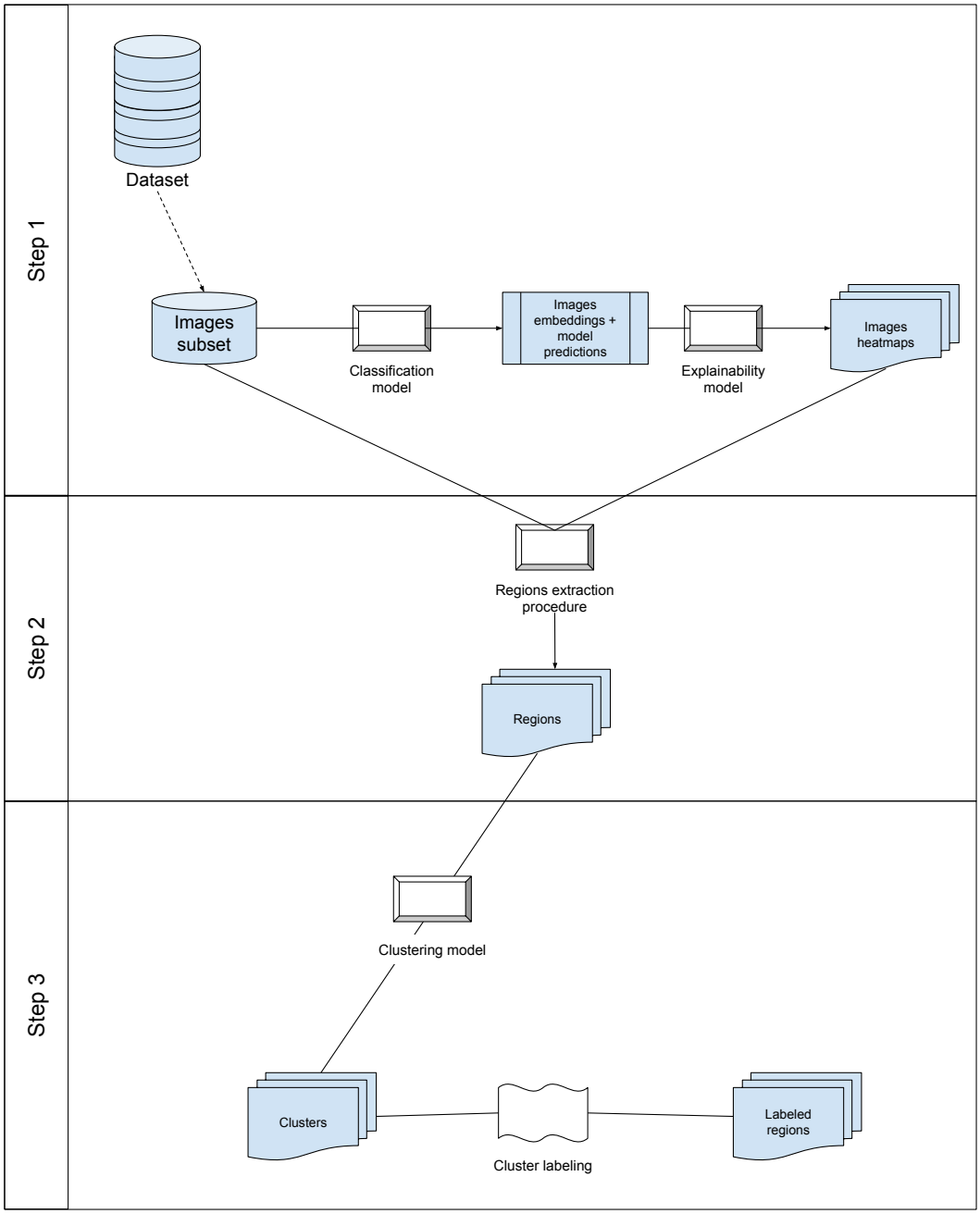


Figure 3.2: Explainability pipeline. The entire process is shown, with the result of the first step being the heatmaps of images, the second step the extracted regions and the last step the labels associated to each extracted region.

3.2.2 Regions extraction

After generating the attribution map, we designed a systematic procedure to extract and save the most salient image regions (*Step 2* in Fig. 3.2). Fig. 3.3 illustrates in detail all steps:

1. **Normalization and preprocessing**: the original input tensor is converted back into an RGB image and normalized to the [0, 1] range. The attribution map is resized to match the resolution of the input image and normalized similarly. If the attribution map has multiple channels, it is converted to grayscale;

2. **Thresholding**: to isolate the most influential pixels, we apply percentile-based thresholding. Specifically, we build a binary mask of the most important regions by keeping only pixels above the 90th percentile;

3. **Morphological processing**: to create clean and contiguous masks, we use a sequence of morphological operations: "**closing**" fills any small gaps or holes, and "**opening**" removes isolated noisy pixels;

4. **Connected components**: the binary mask is decomposed into connected components. For each component, we compute a bounding box; very small regions (with width or height < 10 px) are ignored as noise. Each bounding box is then expanded by 20 px in all directions to capture entire morphological structures;

5. **Region cropping and saving**: the corresponding crop from the normalized RGB image is extracted and saved as a separate region. These regions represent candidate salient biological traits that will later be clustered and labeled.

3.2.3 Clustering phase

After generating the heatmaps and extracting the salient regions, the next step is to assign a label to each extracted region. The goal is to determine whether a region corresponds to a meaningful biological structure (and, if so, which one specifically), or if it reflects a spurious focus on irrelevant elements such as background or human hands. This is *Step 3* visualized in Fig. 3.2.

We embedded each region into a semantic feature space using the image encoder of BioCLIP-2 [21], the same one already used in our classification model (see Sec. 3.1). Since the raw embedding space is high-dimensional, we applied Uniform Manifold Approximation and Projection (**UMAP**[49]) to reduce the dimensionality while preserving local neighborhood structures. UMAP works by constructing a high-dimensional graph representation of the data and then optimizing a low-dimensional

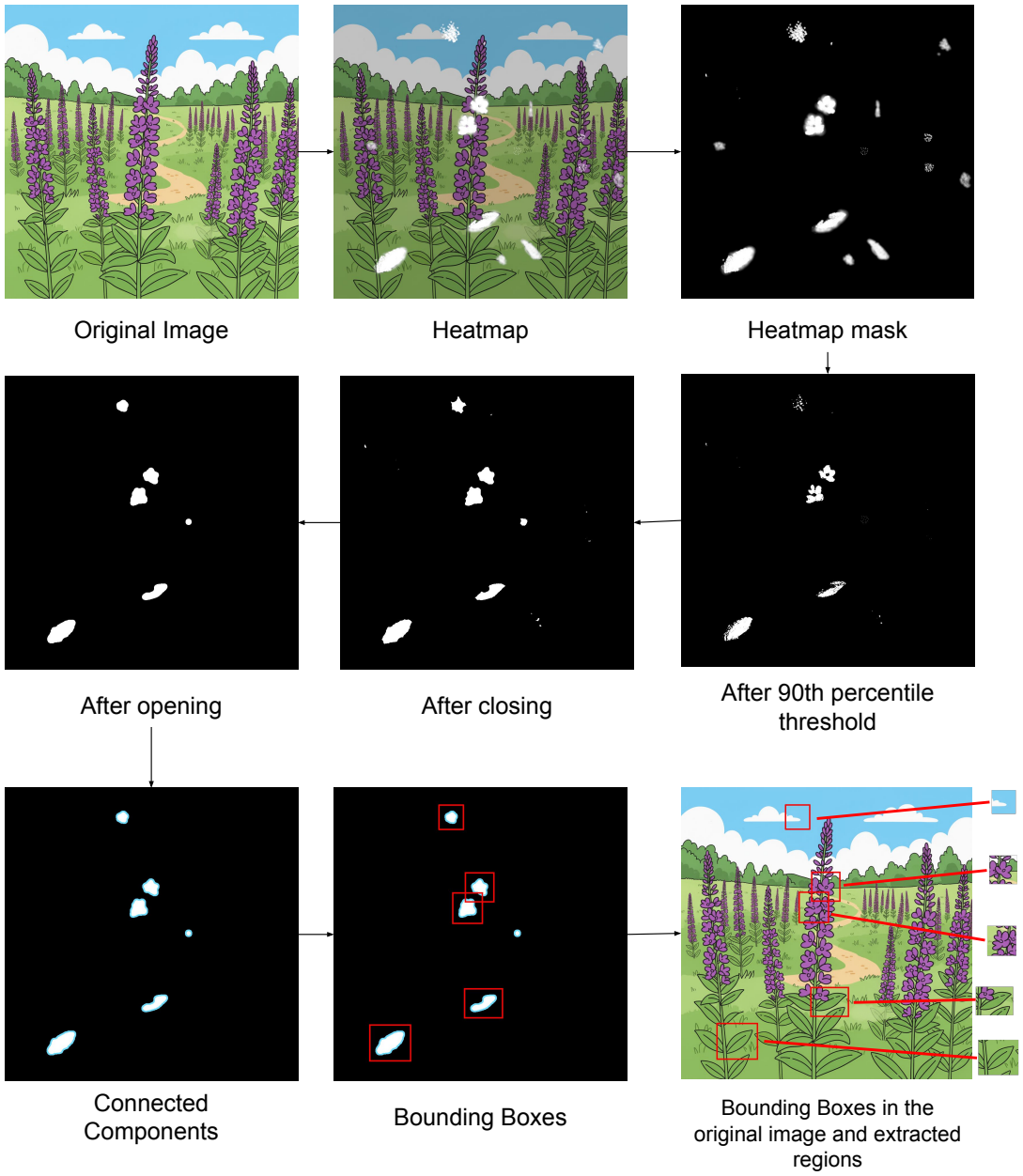


Figure 3.3: Detailed procedure showing how salient regions are extracted from an image and its heatmap.

graph to be as structurally similar as possible, measured using cross-entropy. The reduced embeddings facilitated both visualization and downstream clustering.

We then used a centroid-based clustering method, **KMeans**, an iterative algorithm that repeatedly assigns data to the nearest cluster centroid and then recalculate the centroid's position.

The final output of this stage was a mapping between each extracted region and its assigned cluster.

By visually inspecting the regions within the generated clusters, we manually assigned to each cluster the labels: **Leaf**, **Flower**, **Stem**, **Hand** and **Background/undefined**.

3.3 Final analysis

These preliminary steps enable the final analysis. To extract meaningful patterns from the data, we designed the pipeline illustrated in Fig. 3.4, which consists of three main stages:

1. **Region Extraction**: from the images of the entire dataset, we extract salient regions following the procedure described in Sec. 3.2.3;
2. **Regions Labeling**: the extracted regions are then assigned to clusters using the KMeans model with fixed centroids trained in Sec. 3.2.3. Since clusters were manually annotated in advance, each region inherits the labels of its corresponding cluster.
3. **Pattern Analysis and Discovery**: finally, in Sec. 3.3.1, we analyze the distribution of labeled regions across the dataset in order to identify recurring structures and discover patterns underlying predictions.

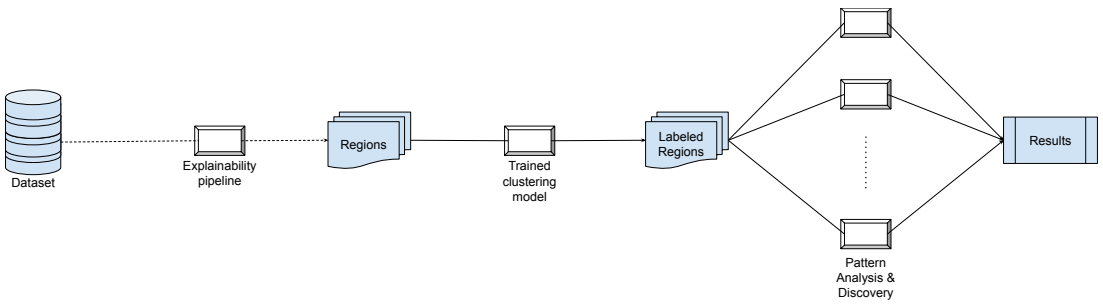


Figure 3.4: Final analysis pipeline. From the dataset we extract regions, we cluster them (using the trained model from Sec. 3.2.3), and label them accordingly. These labeled regions are then used to discover and analyze patterns in the data.

For this final analysis, we expanded the cluster labels: instead of just using *Leaf, Flower, Stem* for plant structures, we included characteristic traits specific to each

species and each plant structure (see Appendix A). *Hand, Backgroung/undefined*, instead, remained the same as previously defined.

3.3.1 Pattern Analysis and Discovery

Region-level outputs were aggregated at the image level. For each image, we computed:

- **Ground truth** and **predicted class** computed by our classification model;
- **Prediction correctness**: binary indicator (if ground truth and predicted class correspond) and error type:
  - **True Positive (TP)** = Invasive plants correctly predicted;
  - **True Negative (TN)** = Non Invasive plants correctly predicted;
  - **False Positive (FP)**= Non Invasive plants predicted as Invasive;
  - **False Negative (FN)** = Invasive plants predicted as Non Invasive.
- **Regions metrics**: number of regions extracted from the image, *Hand* and *Background/undefined* fractions (representing the proportions of regions that contain these labels), coverage fraction (computed as the sum of the areas covered by the regions divided by the total area of the image);
- **Trait metrics**: presence of each characteristic trait (listed in Appendix A), relative frequency within the image, total count of distinct traits present (richness), and the Pielou evenness index, defined as

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

where $H' = -\sum_{i=1}^{S} p_i \log p_i$ is the Shannon diversity of the trait distribution, p_i is the relative frequency of trait i in the image, and S is the number of distinct traits (richness). The index J' ranges from 0 (uneven distribution, dominated by a few traits) to 1 (perfectly even distribution across all traits).

3.4 Experimental settings

3.4.1 Classification Model

In this study, we evaluated three models as image feature extractors for our dataset:

- **ResNet18**: a convolutional neural network (CNN) pretrained on ImageNet, used here as a baseline due to its strong and well-established performance in generic image recognition tasks [50].

- **BioCLIP-1**: a contrastive learning vision model trained on TreeOfLife-10M, a large scale collection of biological images covering plants, animals, and fungi [20].
- **BioCLIP-2**: an improved version of BioCLIP-1 that uses a larger vision transformer and a broader dataset, TreeOfLife-200M, which extends the original collection to a much greater scale and diversity [21].

The three models were compared as embedding extractors. Each model was used to generate a multidimensional embedding for every image in the dataset, which was then employed to train a classifier. For ResNet18 (imported from `torchvision`), the final fully connected layer was removed, while BioCLIP 1 and BioCLIP 2 were imported from HuggingFace and only their image encoders were used. These embedding extractors were not fine-tuned on the dataset, so we evaluated solely their ability to produce meaningful representations without additional training.

After generating embeddings, we divided the dataset into 80% for training and 20% for validation. The distribution of samples across both splits is reported 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 and validation sets for invasive and non-invasive species

To compare the models, we trained the classifier described in Sec. 3.1 for 50 epochs using embeddings extracted by each model. After every epoch, the classifier was evaluated on the validation data. Early stopping was enabled, with a patience of 20 epochs and no minimum improvement threshold, to prevent overfitting and to stop training once the validation loss no longer improved.

The main training parameters are summarized in Tab. 3.3.

Model performance was assessed by the classifier's accuracy after the final training epoch. Results are presented in Tab. 3.4. As expected, BioCLIP-2 achieved the highest performance and already satisfying results, despite no fine-tuning and a relatively short training phase. This confirms its suitability as a feature extractor for our task (see Fig. 3.5).

3.4.2 Explainability and Clustering phase

For both explainability methods (Integrated Gradients and Gradient SHAP) we use the implementation from the `Captum` library; while for UMAP we use its

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 inverse of the logarithms of the class samples, to counter the slight imbalance in the class distribution.

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 taken into consideration, tested as feature extractors from the images. The values for the different metrics report the scores obtained after the last epoch of training.

homonymous library. For Integrated Gradients, we approximate the path integral using 100 interpolation steps with an internal batch size of 5. For both Integrated Gradients and Gradients SHAP, we set the baseline to a completely black image, and the resulting attributions were aggregated across the RGB channels to obtain a single normalized heatmap.

To identify the most suitable explainability method for our pipeline, as well as the optimal hyperparameters for **KMeans** and **UMAP**, we systematically explored different combinations of settings. The range of values considered for each hyperparameter is reported in Tab. 3.5. The quality of each configuration was assessed using the silhouette score computed on the final clustering assignments, as it provides a standard measure of both intra-cluster cohesion and inter-cluster separation.

By combining all the hyperparameter values, we obtained a total of 900 possible configurations for each explainability method, resulting in 1800 experiments overall. The clustering was performed on extracted regions from a subset of 2000 images: Integrated Gradients produced 7509 regions, while Gradient SHAP generated 20021 regions. We observed that a small number of UMAP neighbors (5-10) and low *min_dist* values (0.01-0.025) tend to produce higher silhouette scores. Both

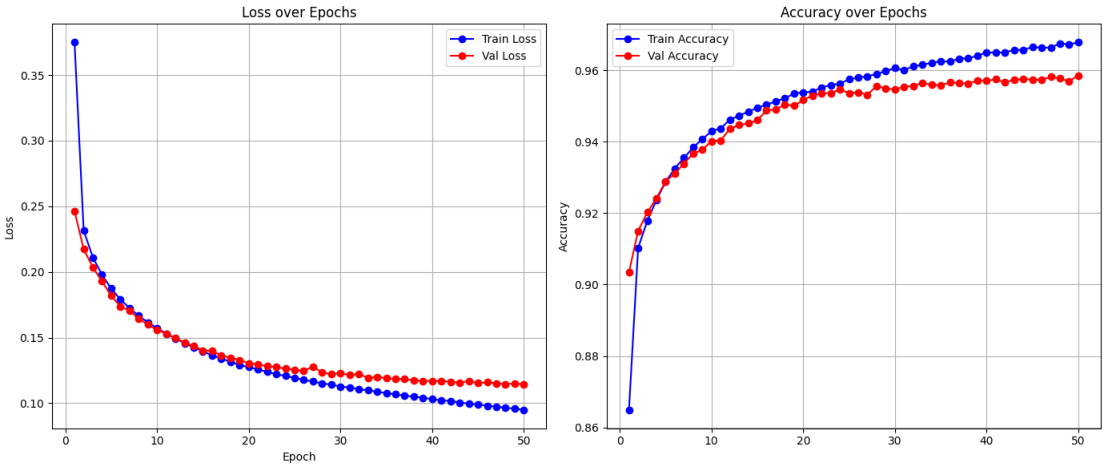


Figure 3.5: Validation loss and validation accuracy throughout the training of the classifier using BioCLIP-2 embeddings. Both metrics have not yet reached a plateau, suggesting that improvements on the results are possible with further training.

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

Table 3.5: Hyperparameters explored for the clustering pipeline.

explainability methods achieved similar silhouette scores, although Integrated Gradients generally resulted in slightly higher values. The configuration with the highest silhouette score was selected for subsequent analysis and is reported in Tab. 3.6: its results are shown 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.

3.4.3 Final Analysis

Predictive Feature Analysis

We applied a Random Forest classifier to identify which image-level features are associated with prediction correctness. The input features included both trait presence and frequency, the number of extracted regions, the number of distinct traits (richness), the coverage fraction, the Pielou evenness index, the hand and background/undefined fractions. Two complementary analyses were performed:

1. a **global** analysis, ranking predictors of correct vs incorrect classifications, and

2. a **error-type** analysis, conducted in a one-vs-all manner (TP, TN, FP, FN) to identify features contributing to specific error patterns.

Feature importance was quantified in two ways. First, *impurity-based importance* measured each feature's average contribution to reducing node impurity across the forest, providing a fast estimate. However, it can be biased toward variables with many categories or continuous scales, and it may overestimate the importance of features that produce strong splits for a small subset of samples. Second, *permutation-based importance* assessed the decrease in forest accuracy when the values of a single feature were randomly shuffled, offering a measure of each feature's true predictive influence that is less susceptible to biases from feature's scale or cardinality.

Metric-specific correlation with accuracy

To further investigate the Random Forest analysis, we analyzed the correlation between specific image metrics and prediction correctness. Each metric was binned into discrete categories, chosen either by quantiles (for continuous metrics) or fixed intervals (for counts), and the average prediction accuracy was computed per bin.

Within each bin, species composition was analyzed by computing the distribution of taxa and comparing it to the overall dataset using the Kullback-Leibler (KL)

divergence:

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

where $P(i)$ and $Q(i)$ are the proportions of species i in the bin and in the overall dataset, respectively. This analysis allowed us to assess whether bins with distinct accuracy patterns also had broadly comparable species distributions; if a bin's accuracy differed but the distribution was highly skewed, the observed difference might reflect species composition rather than the metric itself.

Differences in accuracy across bins were tested using one-way ANOVA, which evaluates whether the mean accuracy differs significantly between two or more groups (bins). In this context, ANOVA determines whether variation in prediction correctness is actually associated with the binned metric, while accounting for within-bin variance.

This approach allowed us to quantify how the chosen metrics are associated with model performance while controlling for potential confounding effects of species composition, completing the feature importance analysis from the Random Forests.

Pairwise Trait Importance and Masked Image Analysis

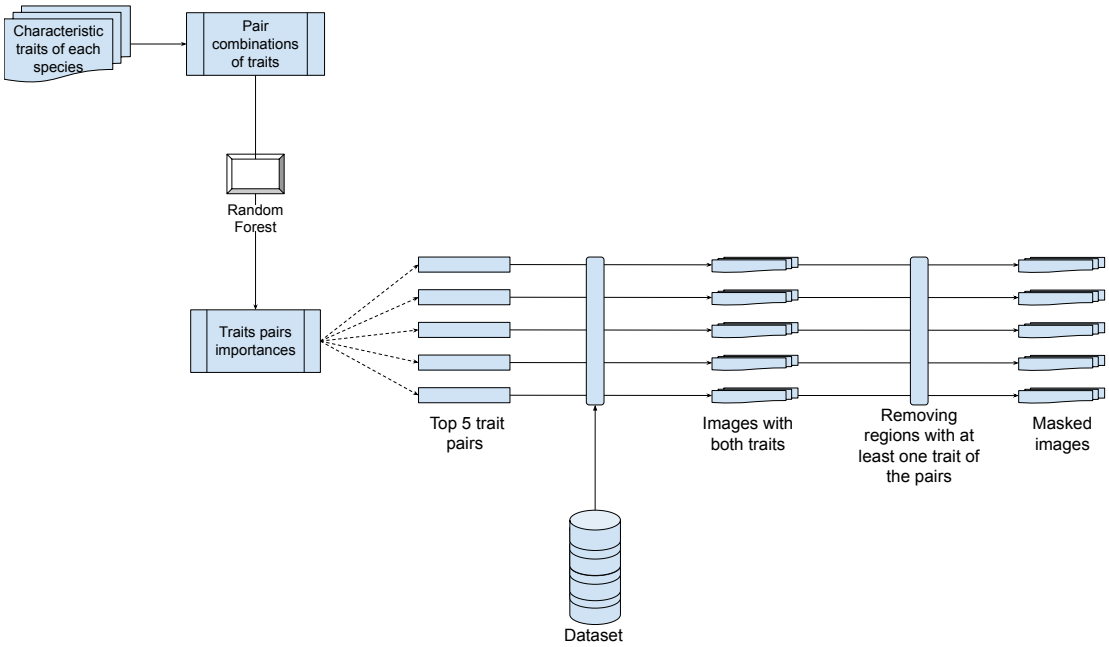


Figure 3.6: Creation of masked images: after identifying the top 5 most important pairs of traits, for each pair, images having both traits are selected. Within these images, regions labeled with at least one of the traits are masked.

To investigate which combinations of traits most strongly influenced model predictions, we performed a **pairwise predictive feature analysis** at the image level. The process is shown in Fig. 3.6.

First, all possible combinations of characteristic traits (defined per species and per plant structure; see Appendix A) were generated. These combinations were grouped into three categories based on their taxonomic specificity: **Common** (traits shared by both invasive and non invasive species), **Non-invasive only**, and **Invasive only**.

Random Forest classifiers were trained using these pairwise features to predict whether the classification was correct. Again, feature importances were computed globally and separately for each error type (True Positive, True Negative, False Positive, False Negative), allowing us to identify the most predictive pairs of traits that contributed to model success or failure. The top five pairs were retained for further analysis.

For each selected pair, all images containing both traits were identified, and the corresponding segmented regions (labeled with at least one of the traits) were localized within the original images. These regions were then masked out (replaced by transparent holes): the resulting 'masked' images were used as inputs for a new round of classification, with the goal of quantifying how the model's prediction accuracy and confidence changed when the most predictive traits were removed.

Comparing model behavior between pairs labeled as *Common* and *Non-invasive only* (no *Invasive only* pairs were available in the dataset) allowed us to assess whether predictions relied more heavily on general morphological patterns shared across taxa or on features distinctive of non-invasive species.

Chapter 4

Results

4.1 Dataset construction

To assemble a dataset of species within the same genus, we first selected a species of interest. We chose purple loosestrife (*Lythrum salicaria*), a member of the family Lythraceae and listed among the world's 100 most invasive species [51].

Our analysis focused on the 40 species in the *Lythrum* genus [52], identifying three as invasive: *Lythrum salicaria* [53, 54], *Lythrum hyssopifolia* [55, 56] and *Lythrum virgatum* [57, 58].

Image data for all species were obtained from `iNaturalist.org`, a large citizen-science platform where users upload photographs of organisms observed worldwide. The iNaturalist API provides several access points. We began by querying the API to match each species name to its internal iNaturalist taxon ID. Species without a valid taxon ID (meaning they were absent from iNaturalist) were removed from consideration. Similarly, species present in the database but lacking any public images were excluded from the dataset.

Using the retrieved taxon IDs, we then downloaded all available images for the remaining species (as of June 20, 2025 [59]). This process yielded 30 species with usable data. The full distribution of images per species is shown in Fig. 4.1. Representative samples from the dataset are presented in Fig. 4.2.

4.2 Classification model

4.2.1 Leave-One-Species-Out Cross-Validation

To assess the accuracy and robustness of the classifier in Sec. 3.1, we applied a Leave-One-Species-Out (LOSO) cross-validation scheme (Fig. 4.3). In this procedure, 30 separate models are trained, one for each species included in the dataset. For each

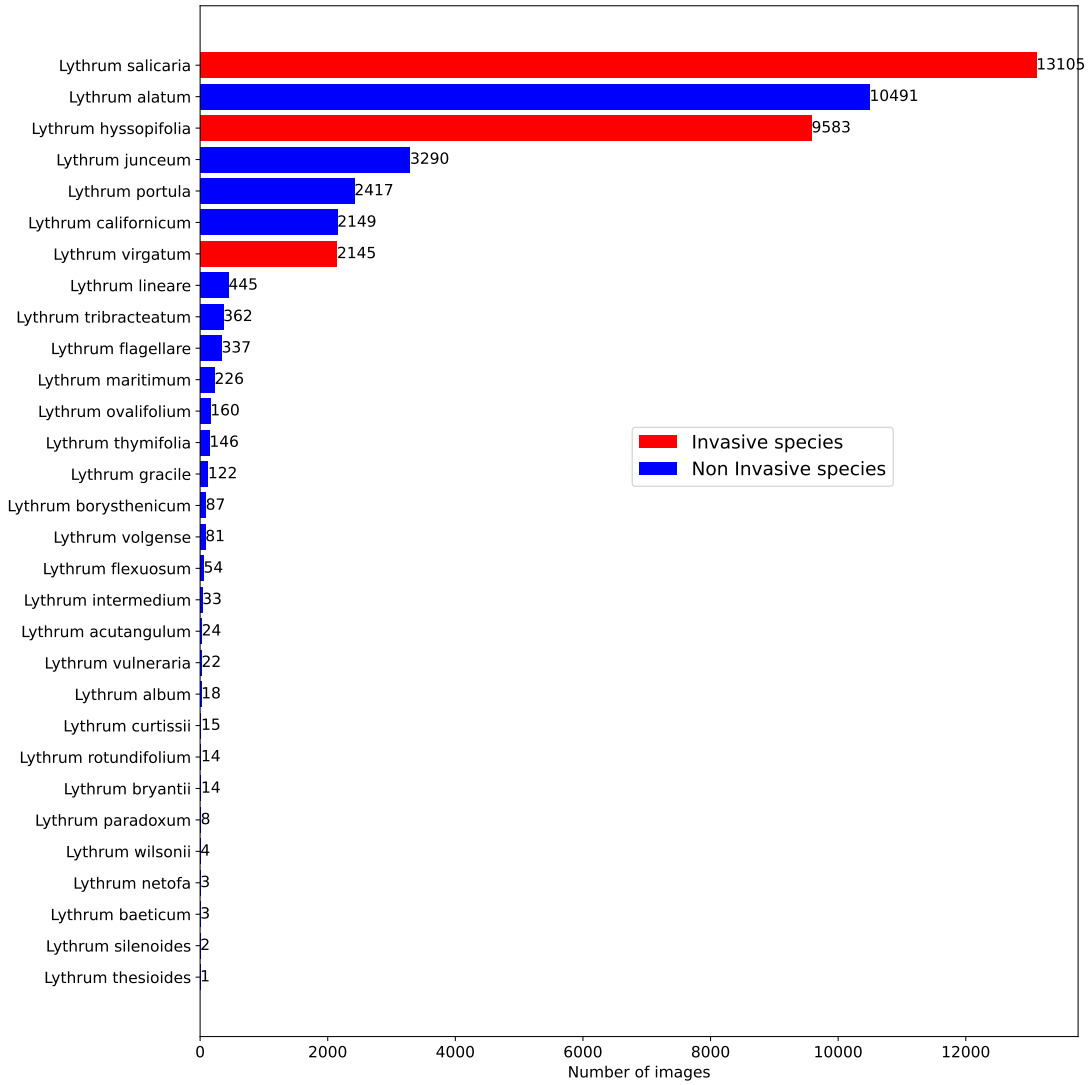


Figure 4.1: The distribution of the total number of images retrieved for each species. The invasive species are shown as a red column whereas the non-invasive species are shown in blue.

(a) **(b)**

(c) **(d)**

Figure 4.2: Examples of images representing different species in the dataset, obtained from iNaturalist: (a) and (b) show invasive species (*Lythrum salicaria* and *Lythrum hyssopifolia*, respectively), (c) and (d) show non-invasive species (*Lythrum ovalifolium* and *Lythrum album*, respectively).

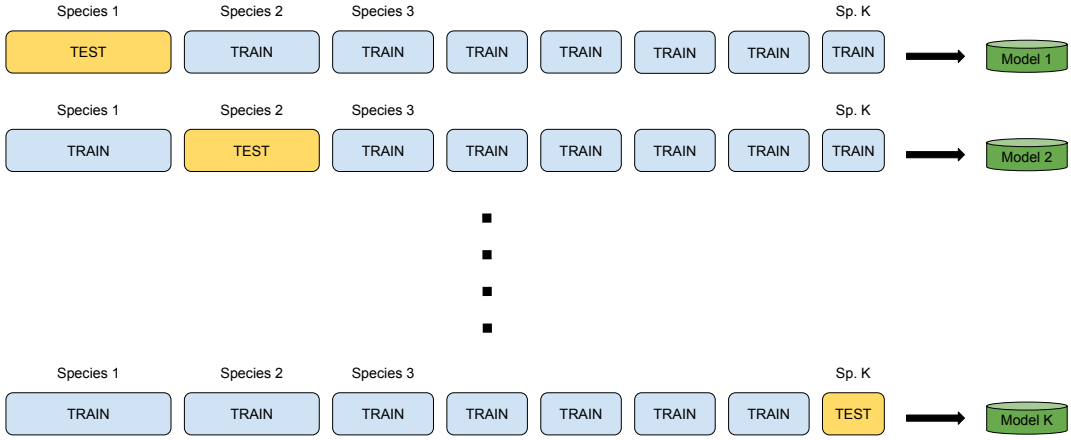


Figure 4.3: Leave One Species Out Cross Validation scheme. Each iteration produces a model tested on a certain species, and is trained on the entire dataset except for that species. Unlike K-Fold Cross Validation, each fold here represents a species, therefore they are not equivalent in sample size (see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

iteration, all images except those belonging to a single species are used for training, while the excluded species serves as the validation set.

This approach allows us to examine the behavior of the model when it encounters a species that was not included in the training phase, relying solely on the morphological traits of other species in the same genus to make predictions.

The results of the Leave One Species Out (LOSO) Cross Validation are reported in Tab. 4.1. These results differ from the accuracy obtained by the classifier trained on all species using an 80-20 split (Tab. 3.4). The average accuracy under LOSO is lower than the overall accuracy of the model trained on all species together, and the standard deviation indicates variability in the results. Several species show moderate accuracy (*L. salicaria, L. junceum, L. virgatum, L. tribracteatum*), while a few have very low accuracy (*L. thymifolia* and *L. wilsonii*, although the small sample size for *L. wilsonii* may limit the significance of this result). For two species (*L. hyssopifolia, L. intermedium*), the model is almost entirely unable to classify them correctly.

These observations suggest that when the model is trained on all species simultaneously, it may be learning to identify the taxon and its classification rather than the traits that determine invasiveness. In contrast, LOSO Cross Validation forces the model to rely on morphological features for predicting invasiveness, which often results in low or unsatisfactory performance when presented with a species it has never seen during training.

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.

4.2.2 Two-Dimensional Mapping of Species Embeddings

To explore how different species are represented in the embedding space, we applied Uniform Manifold Approximation and Projection (**UMAP**) to the image embeddings generated by BioCLIP-2. This step reduced the high-dimensional

vectors to a two-dimensional map (Fig. 4.4), allowing a clearer visual comparison among species. The parameters used for UMAP are listed in Tab. 4.2.

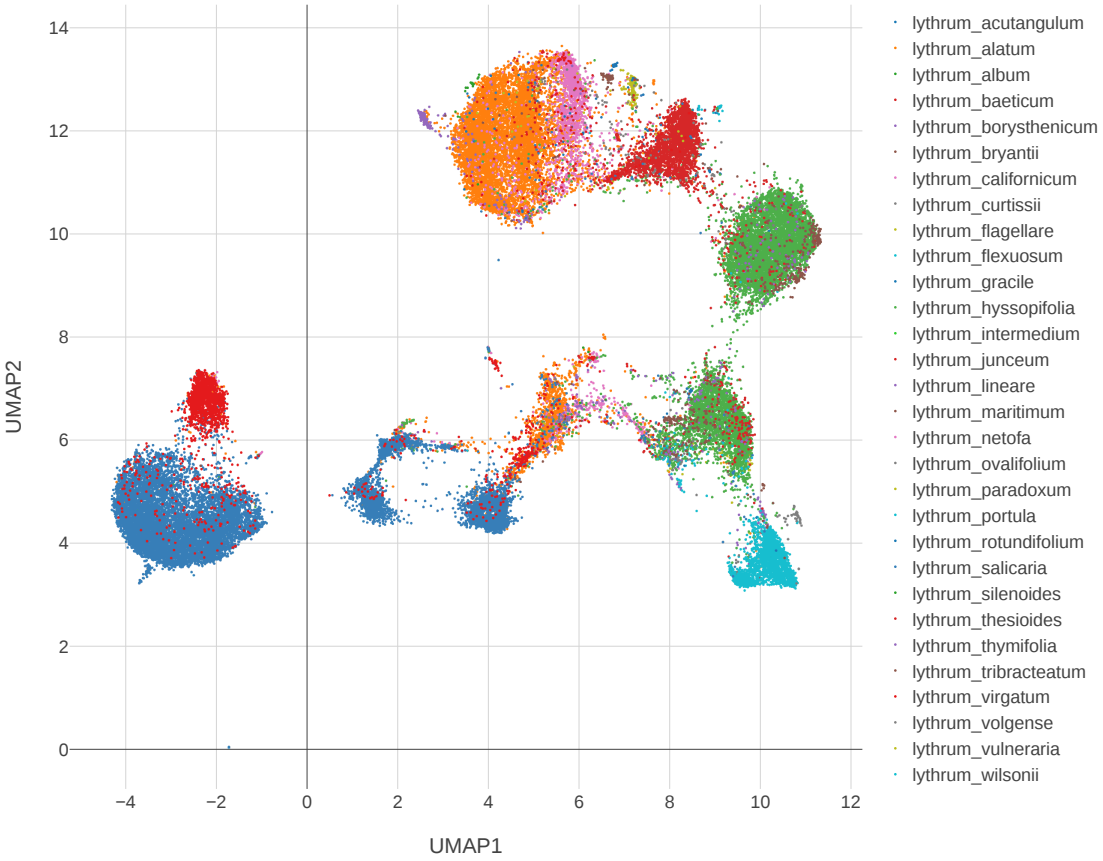


Figure 4.4: Two-dimensional UMAP projection of image embeddings. The invasive species are represented as follows: *Lythrum salicaria* (blue) occupies the region between UMAP1: −5 to 5 and UMAP2: 3 to 6.5; *Lythrum virgatum* (red) spans UMAP1: −5 to 6 and UMAP2: 4 to 8; *Lythrum hyssopifolia* (green)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

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.

We then measured the average distance between species by calculating the mean pairwise distance between their projected points in this 2D space. This provided an estimate of visual similarity and divergence. To keep comparisons reliable, species with fewer than 15 samples were excluded. Across all valid species pairs, the mean distance was 6.584 ± 3.073.

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: Closest (top) and most distant (bottom) species relative to *L. salicaria*, computed using average BioCLIP-2 image embeddings projected in the UMAP space. Only species with at least 15 samples were considered. Invasive species are indicated with (I).

Tab. 4.3 shows the results for *Lythrum salicaria*, the most represented invasive species in our dataset. Its closest neighbor is the non-invasive *Lythrum intermedium*, followed by an invasive species, *Lythrum virgatum*. This proximity suggests that species within the same invasiveness class may share visual traits.

A different pattern emerges for *Lythrum hyssopifolia*, also invasive and well represented in the dataset (Fig. 4.1). Surprisingly, it ranks among the most distant species in the 2D embedding space. The distances listed in Tab. 4.4 show that its five nearest neighbors are all non-invasive, while the other two invasive species occupy the second and third furthest positions. The farthest species overall is again *Lythrum intermedium*.

Finally, for *Lythrum intermedium* (Tab. 4.5), the two invasive species *L. salicaria* and *L. virgatum* appear as its nearest neighbors. The five most distant species are mostly non-invasive, except for *L. hyssopifolia*. The mean distance from *Lythrum intermedium* to the rest of the species is 11.332 ± 2.934.

Some taxonomic sources, as World Flora Online database [60], list *Lythrum intermedium* as a subspecies of *Lythrum salicaria*. This classification helps explain their close spatial proximity in the embedding map and the classifier's inability to correctly identify *Lythrum intermedium* as non-invasive.

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: Closest (top) and most distant (bottom) species relative to *L. hyssopifolia*, computed using average BioCLIP-2 image embeddings projected in the UMAP space. Only species with at least 15 samples were considered. Invasive species are indicated with (I).

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: Closest (top) and most distant (bottom) species relative to *L. intermedium*, computed using average BioCLIP-2 image embeddings projected in the UMAP space. Only species with at least 15 samples were considered. Invasive species are indicated with (I).

4.2.3 Lythrum hyssopifolia exclusion

The results in Tab. 4.1, together with the distances shown in Tabs. 4.3 and 4.4, led us to reconsider the role of *Lythrum hyssopifolia* in the dataset.

During the cross-validation experiments, *L. hyssopifolia* reached an accuracy close to zero. This poor performance likely stems from its large visual distance from other invasive species and its proximity to non-invasive ones (Tab. 4.4). Likewise,

Lythrum salicaria and *Lythrum virgatum* showed only moderate accuracy. Their separation from *L. hyssopifolia* may have reduced their performance, though their mutual closeness (Tab. 4.3) partially offsets this effect. This pattern supports the idea that invasive species share visual traits, while distant embeddings can weaken classification.

To verify this, we repeated the LOSO cross-validation, this time excluding *L. hyssopifolia* entirely. The modified experiment used 29 folds instead of 30.

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).

As shown in Tab. 4.6, removing this species led to a notable rise in mean accuracy across folds and a clear reduction in result variability.

Examining individual species, *L. salicaria* shows a 13.8% increase in accuracy, while *L. virgatum* remains consistent. Results for both remain mediocre but above random chance, partly due to dataset imbalance: after removing *L. hyssopifolia*, there are 15,150 'invasive' and 20,528 'non-invasive' samples. For LOSO Cross Validation, folds for *L. salicaria* and *L. virgatum* are highly unbalanced in the 'invasive' class, which should be considered when interpreting results.

The largest gains in accuracy are observed for *L. junceum, L. tribracteatum*, and *L. thymifolia* (*L. wilsonii* has too few samples to allow for a meaningful analysis), which are the three species most closely related to *L. hyssopifolia*, and all of them are non-invasive.

Overall, these results strenghten confidence in BioCLIP-2 as a reliable embedding extractor and validate our training methodology. We therefore treat *Lythrum hyssopifolia* as a potential outlier and exclude it from the dataset.

We decided to keep the 29 models (each trained by excluding a single species from the training set) separate for the subsequent analyses and experiments. This approach ensures completely unbiased results, as each species is evaluated using a model that has never encountered it during training.

4.3 Explainability pipeline

In Sec. 3.2 we described the explainability pipeline (see Fig. 3.2). In this section, the results of this process are described.

4.3.1 Heatmap generation

To illustrate the outcome of the explainability methods (*Step 1* of Fig. 3.2), Fig. 4.5 presents, for selected samples, the original image alongside the computed heatmap and the resulting overlay highlighting the most influential regions. As illustrated in the examples, in some cases the attribution maps align well with biologically meaningful traits, such as specific plant structures (Figs. 4.5a to 4.5d). However, in other cases, the highlighted regions do not correspond to relevant biological features: for instance, in Fig. 4.5e the method emphasizes a human hand present in the image, while in Fig. 4.5f the attribution is diffuse and does not clearly point to any identifiable structure.

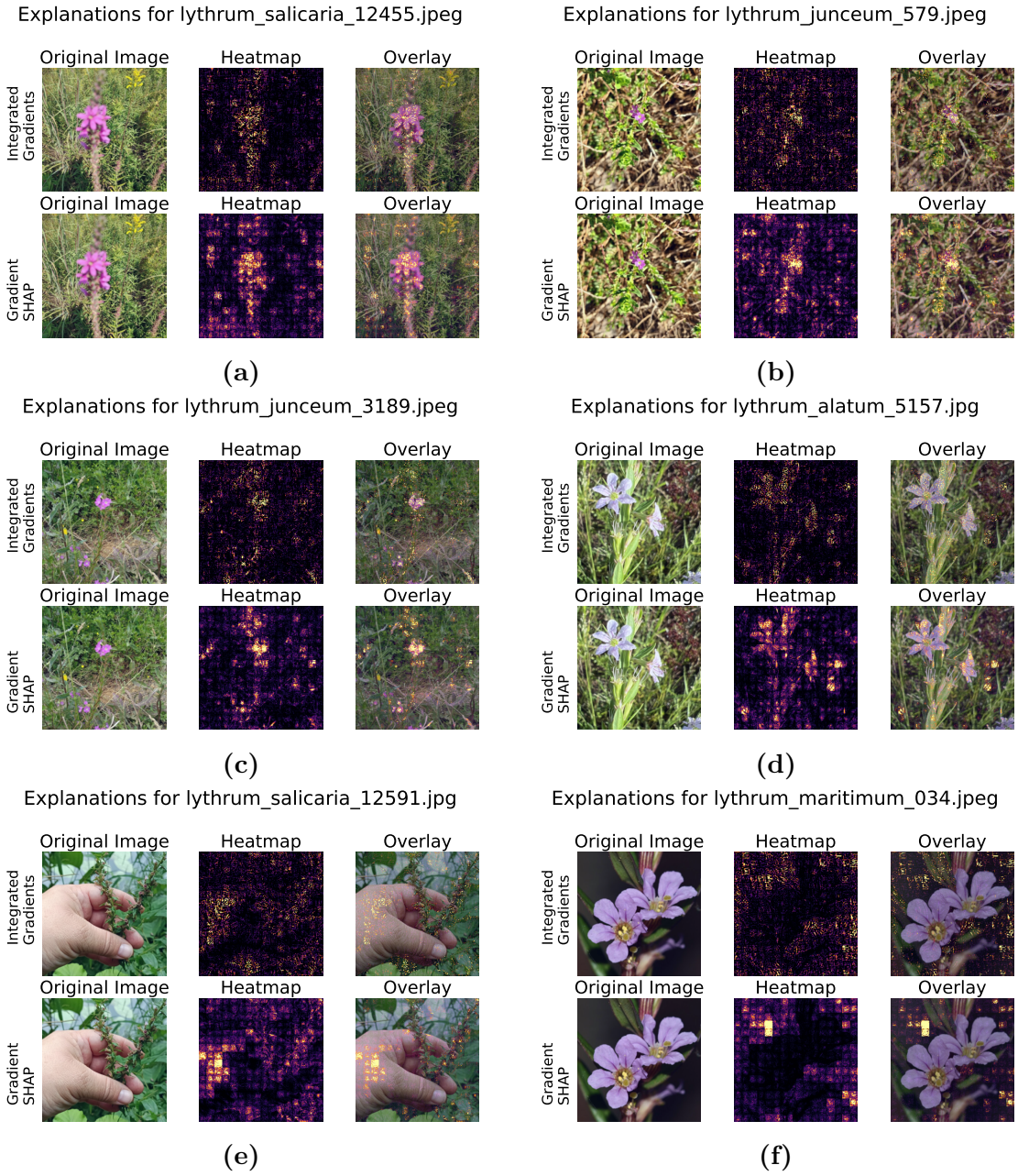


Figure 4.5: Examples of generated heatmaps. For each sample, the original image, the heatmap created and the overlay between the original image and the heatmap are displayed from left to right. Both Integrated Gradients (top row) and Gradient SHAP (bottom row) results are displayed for each sample. In most samples (a-d), the highlighted regions correspond to meaningful biological structures, whereas in some cases (e-f) they do not align with expected features.

4.3.2 Regions extraction

For the first two samples shown in Fig. 4.5, we illustrate the outcome of the region extraction process (*Step 2* in Fig. 3.2) in Figs. 4.6 and 4.7. We can visually note that Gradient SHAP (Figs. 4.6a and 4.7a) systematically produces a larger set of patches compared to Integrated Gradients (Figs. 4.6b and 4.7b). For instance, in Sec. 3.4.2 we already observed that from a sample of 2000 images randomly taken from our dataset, Gradient SHAP extracts 20021 regions while Integrated Gradients produced only 7509. For both methods, the extracted regions naturally vary in size (and consequently in resolution), as expected from their creation pipeline.

From a visual inspection, it is not clear whether Integrated Gradients or Gradient SHAP performs better within our pipeline. The final choice of the most suitable explainability technique was guided by the analysis performed in the clustering hyperparameters search (see Sec. 3.4.2): as shown in Tab. 3.6, **Integrated Gradients** is chosen.

(a)

(b)

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

(a)

(b)

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

4.3.3 Clustering phase

Sec. 3.2.3 described the implementation of the clustering phase (*Step 3* in Fig. 3.2): in this section we analyze its results; parameters used here are shown in Tab. 3.6.

Clustering was first performed on a representative subset of 2000 images from our dataset. This allowed us to manually inspect the resulting clusters and assign them labels based on visual inspection. Once the clusters were labeled, the same cluster model (with fixed centroids) was applied to the entire dataset in the final analysis (see Sec. 3.3). Selecting Integrated Gradients as the explainability method yields a total of 7509 extracted regions, which are then partitioned in 30 clusters (see Tab. 3.6).

We employed **MiniBatchKMeans**: this is a scalable variant that updates the centroids using small random batches of the data rather than the entire dataset, significantly reducing computational cost while retaining similar clustering quality.

To analyze the clustering result, we provide a set of complementary visualizations. 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 reports the distribution of cluster sizes, highlighting the overall homogeneity of the partition

and complementing the cluster size entropy values (see Tab. 3.7). Finally, Fig. 4.10 quantifies centroid convergence by showing their average displacement across batches, illustrating that movements progressively decrease and stabilize after a few iterations.

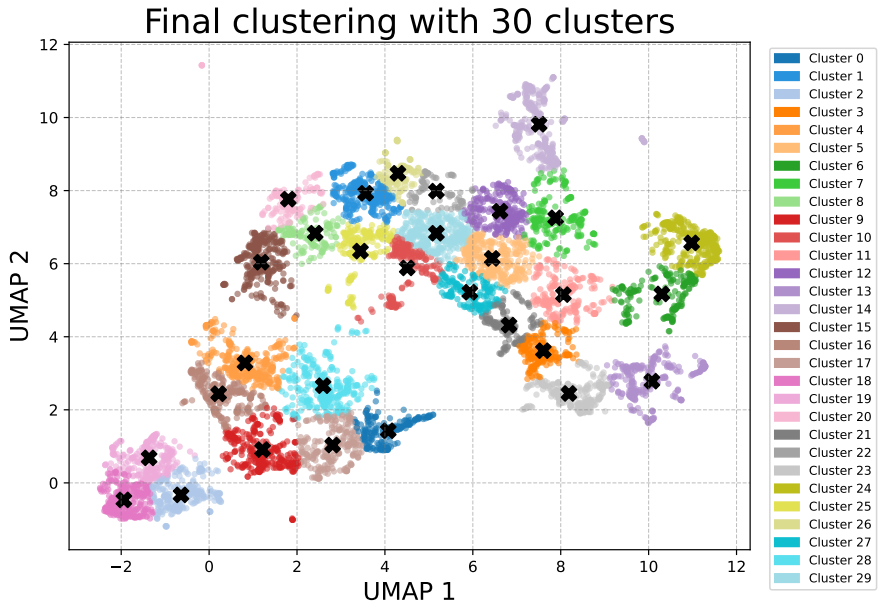


Figure 4.8: UMAP projection of the embeddings with final cluster assignments. Colors denote clusters and black markers indicate the final centroid positions.

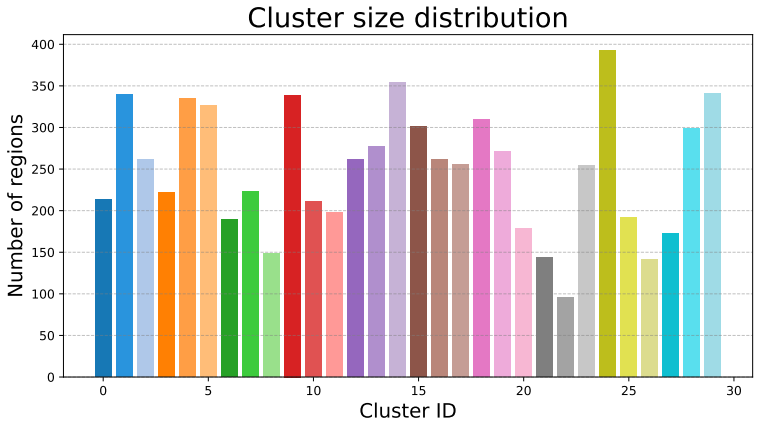


Figure 4.9: Distribution of cluster sizes for the final configuration. Colors match the corresponding clusters in Fig. 4.8.

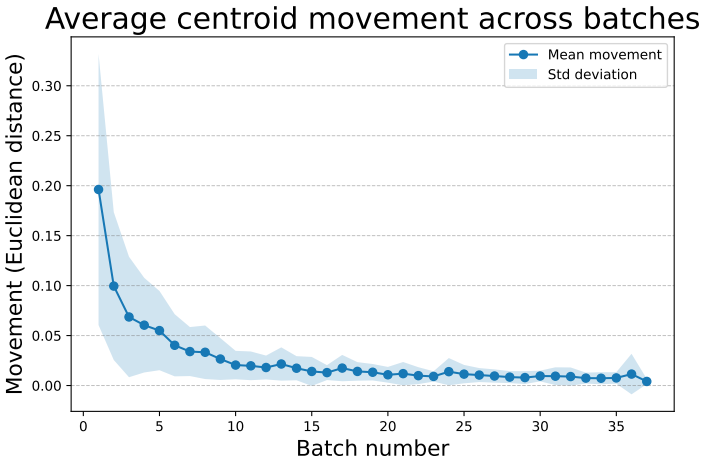


Figure 4.10: Average centroid displacement across minibatch updates. The line shows the mean movement across all centroids, the area represents the standard deviation.

Cluster labeling

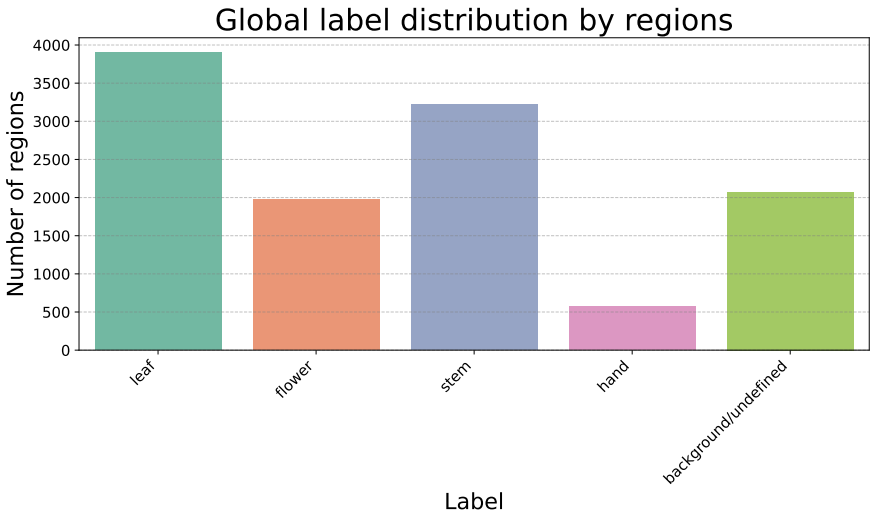
In this phase, clusters were manually annotated through visual inspection of the regions assigned to them. Each cluster received one or more labels describing the dominant visual content. For clarity, we organized the labels into two main categories:

- *Plant structures*:
  - **Leaf**: clusters primarily showing leaves;
  - **Flower**: clusters characterized by the presence of flowers;
  - **Stem**: clusters clearly depicting stems.
- *Spurious or non-informative features*:
  - **Hand**: clusters containing primarily parts of human hands or skin;
  - **Background/undefined**: clusters dominated by background fragments or visually ambiguous regions.

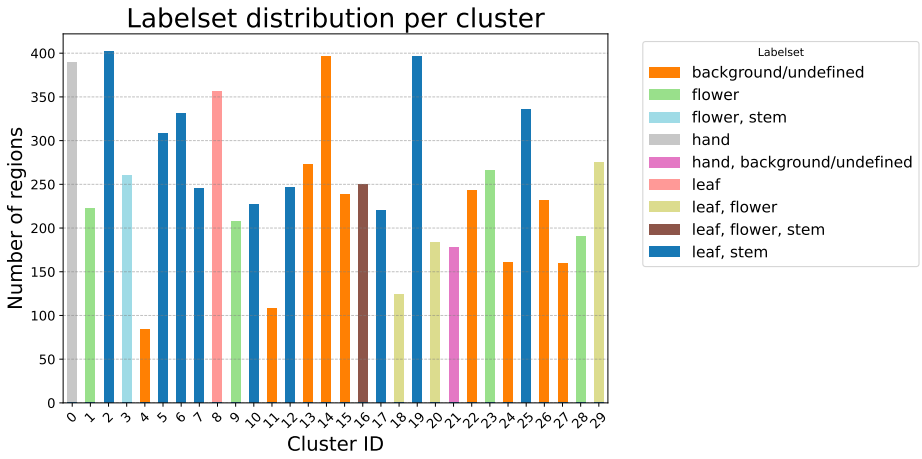
Labels within the same category are not mutually exclusive (e.g., a cluster may simultaneously be labeled *Leaf* and *Flower*), whereas labels from different categories are mutually exclusive. For instance, a cluster cannot be labeled both *Leaf* and *Hand*. This design reflects the aim of our analysis: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

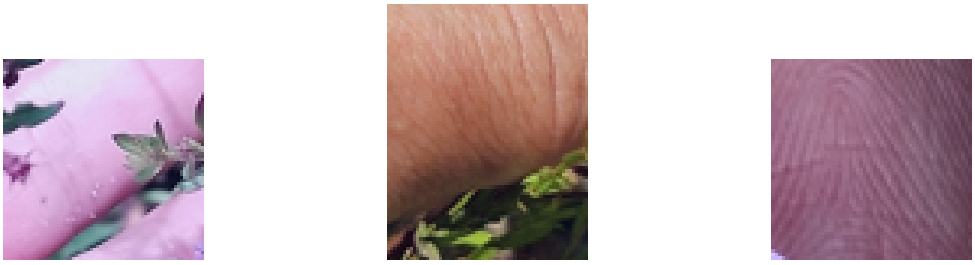


(a)



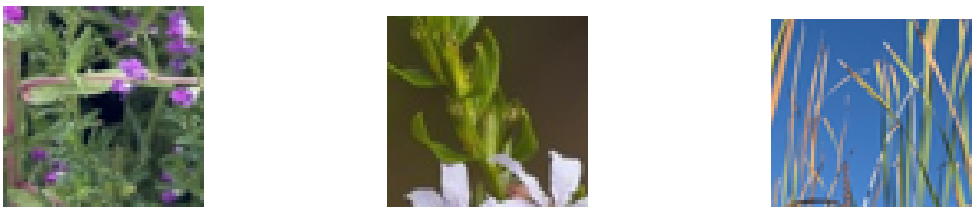
(b)

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



(a) **(b)** **(c)**

Figure 4.12: Three of the regions from the cluster with id=0, which was assigned the label **Hand**. Regions (a) and (c) correspond to *Lythrum alatum*, while region (b) corresponds to *Lythrum californicum*.



(a) *Lythrum junceum* **(b)** *Lythrum alatum* **(c)** *Lythrum salicaria*

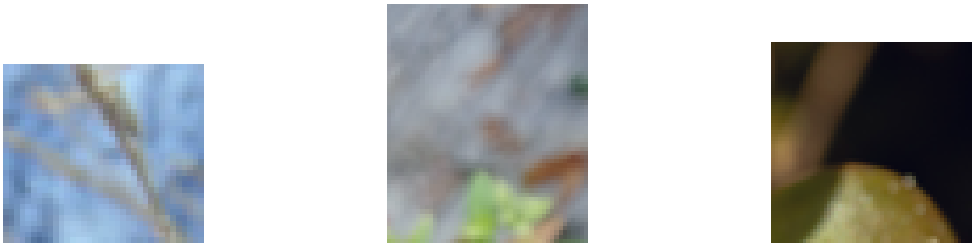
Figure 4.13: Three of the regions From the cluster with id=16, which was assigned the label **Leaf, Flower, Stem**. Region (a) corresponds to *L. junceum*, region (b) to *L. alatum*, and region (c) to *L. salicaria*.

(a) **(b)** **(c)**

Figure 4.14: Three of the regions from the cluster with id=23, which was assigned the label **Flower**. Region (a) corresponds to *L. californicum*, region (b) to *L. junceum* and region (c) to *L. lineare*.

Clustering validation

To validate both the KMeans clustering results and our manual cluster labeling, we applied an alternative density-based clustering method, **HDBSCAN**. Using the same set of extracted regions, we re-cluster the data and analyze how HDBSCAN



(a) **(b)** **(c)**

Figure 4.15: Three of the regions from the cluster with id=27, which was assigned the label **Background/undefined**. Region (a) corresponds to *L. californicum*, region (b) to *L. salicaria*, and region (c) to *L. portula*.

partitions the regions, with particular attention to the consistency of these partitions with our assigned labels. We followed the same pipeline previously adopted for KMeans (see Sec. 3.2.3): the extracted regions were embedded with BioCLIP-2 image encoder, reduced in dimensionality using UMAP, and then clustered. To ensure comparability, we used the regions extracted with Integrated Gradients and used the same UMAP parameters reported in Tab. 3.6. For HDBSCAN the key parameter is the minimum cluster size; to obtain a comprehensive analysis, we explored values in the range

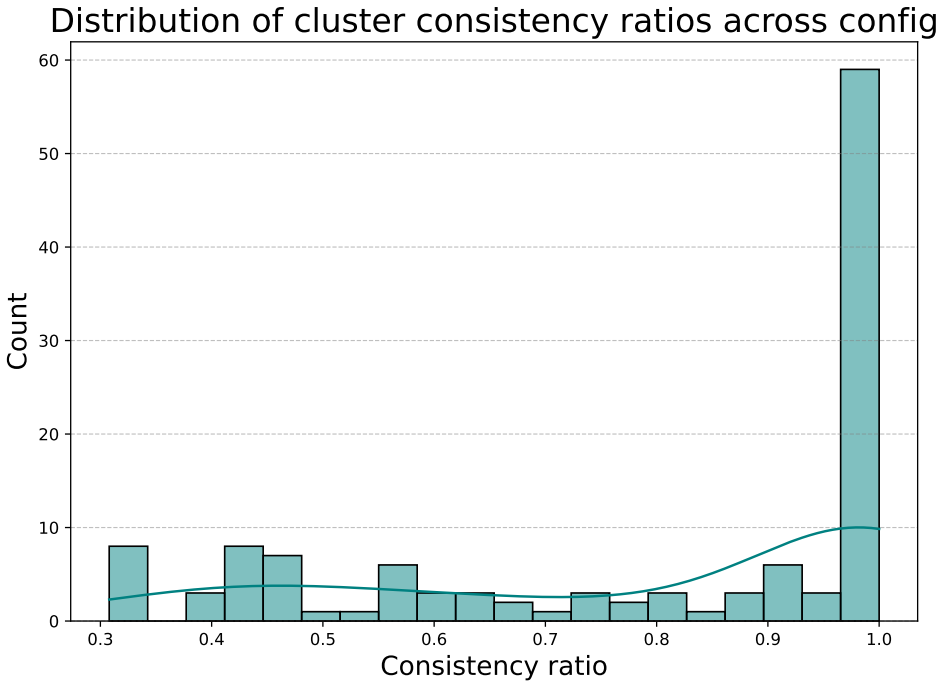
$$min\_cluster\_size = [5, 6, \ldots, 71]$$

.

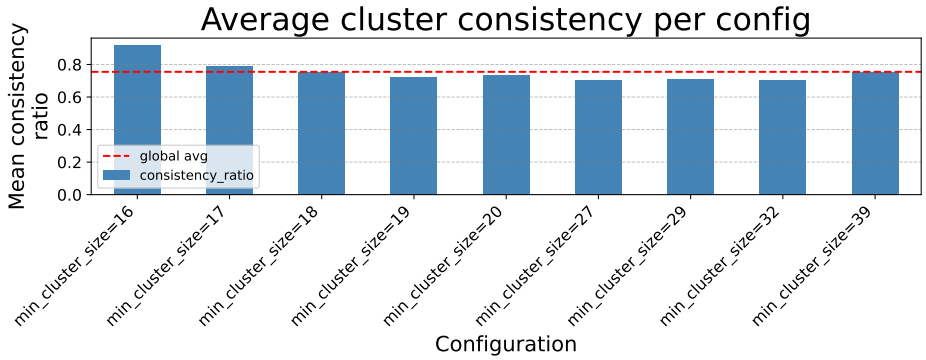
To achieve comparability with the KMeans results, we discarded HDBSCAN configurations that produced fewer than 10 clusters or more than 100 clusters. As a consequence, out of 66 tested configurations, only 9 met this criterion. This outcome is not surprising, as HDBSCAN follows a fundamentally different clustering paradigm: unlike KMeans, which partitions the data into a predefined number of clusters, HDBSCAN identifies clusters of varying density and labels points that do not belong to any dense region as noise. Nevertheless, we consider this reduced set of valid configurations as sufficient for our purposes: since the goal of this step was not to optimize HDBSCAN itself, but rather to validate the consistency of the KMeans results and of our manual labeling, even a modest number of data-driven configurations provides a robust basis for comparison.

Fig. 4.16 summarizes the results: for each valid HDBSCAN configuration, we computed the cluster label consistency ratio, defined as the fraction of images belonging to the dominant label set in each cluster. This metric provides a measure of cluster purity with respect to our manually assigned labels. Across all non-skipped configurations, the average cluster consistency with our manually assigned labels was 0.75, with the best configuration ($min\_cluster\_size = 16$) reaching

0.92 (see Fig. 4.16b). From Fig. 4.16a we can see that the vast majority of clusters across all considered configurations achieve a 1.0 consistency ratio, meaning strong agreement with our pipeline. Appendix B presents a more detailed analysis of our clustering validation.



(a)



(b)

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.

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

4.4 Final analysis

Fig. 3.4 illustrates the pipeline of our final analysis, as described in Sec. 3.3. In this section, we present and analyze the corresponding results.

The dataset consists of 35678 images. From these, using our designed pipeline (Sec. 3.2.2), we extracted a total of 132128 regions. The outcome of the subsequent clustering phase are presented in Fig. 4.17. As shown in Fig. 4.17a, once a region is assigned to a cluster, it inherits the labelset associated with that cluster. Figs. 4.17b and 4.17c display the distribution of labels and labelsets, respectively, for all regions extracted from the dataset.

4.4.1 Predictive Feature Analysis

As described in Sec. 3.4.3, we performed a predictive feature analysis with Random Forests to assess the importance of features (both characteristic traits listed in Appendix A and other metrics computed as described in Sec. 3.3.1). In Tabs. 4.7 to 4.9 results for the global analysis, for True Positive and True Negative analysis, and for False Positive and False Negative analysis are displayed, respectively.

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.

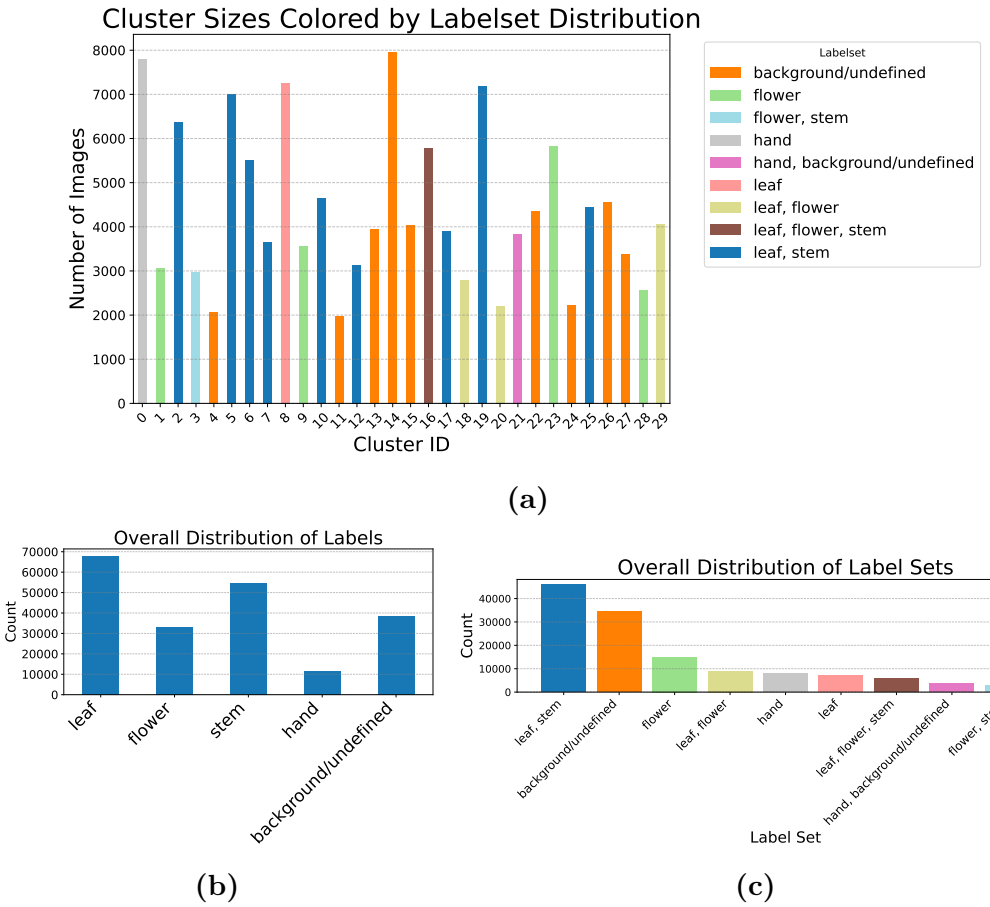


Figure 4.17: Results of the clustering and labeling of regions extracted from the entire dataset. In particular, cluster sizes distribution colored by the corresponding labelset of each cluster (a), labels distribution across all clusters (b) and labelset distribution across all clusters (c) are shown.

Tab. 4.7 summarizes the global feature importance for correct predictions across all images. The fraction of the image covered by extracted regions (`coverage_frac`) is by far the most important predictor (impurity importance= 0.6198; permutation importance= 0.2224), indicating that images with larger region coverage are much more likely to be correctly classified. Secondary features include `pielou_evenness`, `background_frac` and `hand_frac`, which contribute moderately to global prediction accuracy. Permutation importance is generally smaller than impurity, but confirms the rank order of the main predictors. Interestingly, features as `richness` and `n_regions` show low impurity importance but moderate permutation importance, suggesting they might interact with other features to affect model performance.

Overall, the global analysis identifies the strongest single predictors but may obscure feature-specific effects that vary across different types of predictions.

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.

Error-Type Analysis

To capture more detailed relationships, we performed a separate analysis for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Tabs. 4.8 and 4.9). This approach allows us to determine which features drive specific types of correct and incorrect predictions, providing a more nuanced

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.

understanding than the global analysis alone.

For TP and TN outcomes, `coverage_frac` remains the dominant predictor (TP: 0.4099 impurity, 0.1642 permutation; TN: 0.1650 impurity, 0.0881 permutation), confirming its strong influence on correct classification. Traits related to leaf base morphology (`rounded_at_the_base`) and to flower morphology (`flowers_in_whorled_clusters`) also show elevated importance, suggesting that these features help distinguish between true positive and true negative classifications. Permutation importance generally aligns with impurity ranking, supporting

robustness of the identified predictors. Notably, while the overall patterns are similar, TP outcomes show higher reliance on `coverage_frac`, whereas TN outcomes emphasize certain trait-specific features slightly more, highlighting subtle differences in the drivers of correct positive versus correct negative predictions.

For FP and FN outcomes, `coverage_frac` again dominates (FP: 0.6493 impurity, 0.0684 permutation; FN: 0.5372 impurity, 0.1558 permutation). Other contributors include `pielou_evenness`, `background_frac` and `hand_frac`. Permutation importance highlights additional moderate contributors, such as `n_regions` and `attenuate_at_the_base_freq`, suggesting that feature interactions and nonlinear effects influence error patterns. Notably, FP and FN patterns are broadly similar.

4.4.2 Metric-specific correlation with accuracy

Region Coverage

As suggested by the Random Forest analysis, the fraction of the image covered by extracted regions (`coverage_frac`) emerged as the most influential predictor of model correctness. To further investigate this relationship, we analyzed how prediction accuracy varies across different levels of region coverage. Images were grouped into eleven coverage categories, and the mean accuracy, sample size, and corresponding species distribution divergence (KL divergence) were computed for each bin (Tab. 4.10, Fig. 4.18).

Overall, results reveal that classification accuracy remains largely stable across most coverage ranges, fluctuating between 0.76 and 0.79 up to 25% coverage (Fig. 4.18a). Only the smallest ($< 1\%$) and largest ($> 25\%$) bins show deviations, although these categories include very few samples (220 and 65 images respectively), making their estimates less reliable. The extremely high accuracy in the last bin (100%) is based on a single image and therefore not meaningful. Species distribution divergence (KL) remains low across all major bins (< 0.01), confirming that these accuracy patterns are not driven by differences in taxonomic composition (Fig. 4.18b).

These findings indicate that, contrary to what the Random Forest importance might suggest, region coverage alone does not strongly determine prediction accuracy. The Random Forest model likely interpreted the separation between low- and high-coverage samples as a strong discriminative signal, even though the underlying relationship is weak or non-causal. This behavior is consistent with how impurity-based importance in Random Forests operates: features that allow a clear partition of the data, even if only over a small subset of samples, can receive disproportionately high importance. In our case, a few extreme high-coverage samples show higher accuracy and are easily separable from the rest. The model therefore identifies coverage as a useful split criterion, despite its limited general predictive power across the dataset.

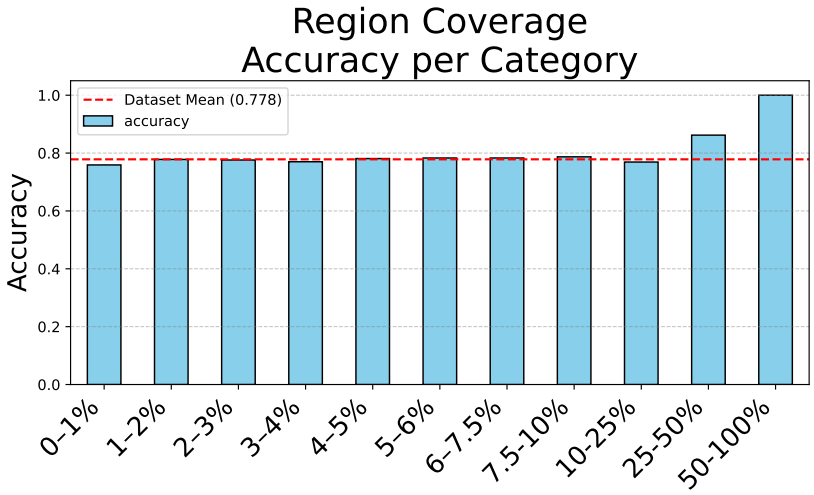
Category	Accuracy	Sample size	KL Divergence
0-1%	0.759	220	0.009
1-2%	0.778	2977	0.007
2-3%	0.776	2134	0.008
3-4%	0.770	3891	0.005
4-5%	0.781	3419	0.003
5-6%	0.783	3754	0.032
6-7.5%	0.783	5086	0.002
7.5-10%	0.787	6035	0.003
10-25%	0.769	6807	0.007
25-50%	0.862	65	0.201
50-100%	1.000	1	1.22

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.

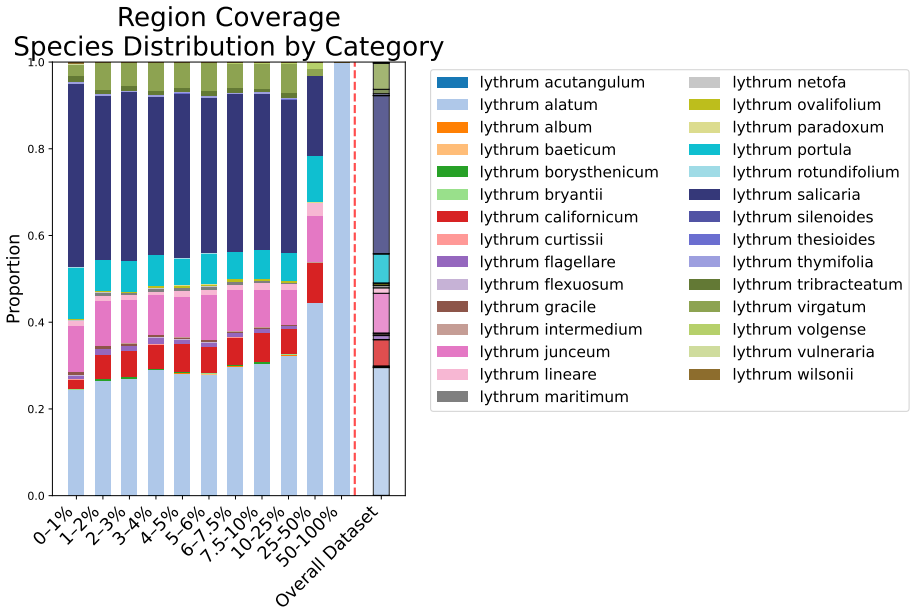
ANOVA: $F = 1.221$, $p = 0.2712$; the highest mean accuracy is observed in the 50-100% coverage category (1.000) and the lowest in the 0-1% category (0.759). No strong statistical evidence of differences across categories.

It is also important to note that the vast majority of images in our dataset exhibit very low coverage, while samples with higher coverage values are comparatively rare and display larger KL divergence from overall species distribution. This indicates that these images are not representative of the general dataset and may bias the Random Forest's assessment of importance. Consequently, within available data, there is **no clear evidence of a consistent dependency between coverage and prediction accuracy**, although such a relationship cannot be completely excluded given the strong imbalance in the coverage distribution.

In addition to region coverage, the same analysis was extended to all other metrics identified as potentially relevant by the Random Forests, 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 reported in Appendix C. Overall, no other metric showed a clear or consistent relationship with prediction accuracy, except for weak or dataset-dependent effects. In particular, while some metrics exhibit statistically significant trends, these patterns were often accompanied by strong differences in species composition across bins (high KL divergence), suggesting that the observed variations are likely driven by taxonomic imbalance rather than intrinsic effects of the metrics themselves.



(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.

4.4.3 Pairwise Trait Importance and Masked Image Analysis

In Sec. 3.4.3 we described a new Random Forest analysis that differs from the previous one by the use of combinations of pairs of traits, here we illustrate its results. The analysis produced feature importance scores for each combination, computed globally and separately for each error type (TP, TN, FP, FN). Tab. 4.11 lists top 15 evaluated pairs and their corresponding importance values across these categories. The results were highly consistent across different Random Forest runs, with only minor variations in exact rankings.

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.

From these results, we selected five pairs of traits for further analysis (Tab. 4.12). Importantly, these were not the top five globally, but rather the top five in terms of importance within the True Positive and True Negative categories, ensuring that the selected pairs are most relevant for correctly predicting both invasive and non-invasive species. The selected pairs belong either to the *Common* category, shared by invasive and non-invasive species, or to the *Non-Invasive only* category, characteristic of non-invasive species. No *Invasive only* pairs were available in the dataset.

For each of the selected trait pairs, we next identified all images containing both traits, along with the corresponding regions within those images that exhibited

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.

at least one of the traits, with the counts summarized in Tab. 4.13. This data represents the subset of images and regions that were targeted for removal in the subsequent masked-image experiments. Examples of these images and the relative traits can be seen in Fig.4.19 and 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%).

To assess how the removal of the most predictive trait combinations affected model behavior, we evaluated classification performance and prediction confidence on the masked images. Tab. 4.14 summarizes the results in terms of overall accuracy, number and rate of label flips (images whose predicted label changed after masking), and the variation in True Positive (TP) and True Negative (TN) counts. Changes

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.

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.

in accuracy are also highlighted for each pair of traits considered in Fig. 4.21.

For pairs belonging to the *Common* category, the impact of masking was minimal. Accuracy decreased by less than 2%, and flip rates remained around 17%. TP and TN counts varied only slightly, indicating that these trait combinations, although frequently present, are not strictly necessary for the classifier to make correct

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.

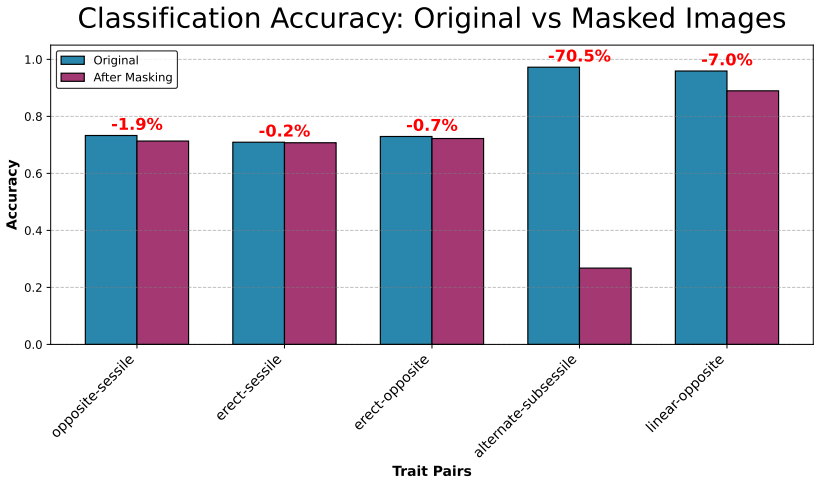


Figure 4.21: For each pair of traits, accuracy computed with original images (in blue) and accuracy computed with masked images (in magenta) are shown; on top of each bar, the difference in accuracy is computed in red.

invasiveness prediction.

In contrast, masking pairs belonging to the *Non-Invasive only* category produced substantially larger effects. In particular, removing regions associated with the pair *Alternate-Subsessile* caused a dramatic drop in accuracy (from 0.97 to 0.27) and led to more than 70% of the images flipping their predicted class. Similarly, although to a lesser extent, masking *Linear-Opposite* reduced accuracy by 7%.

For 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 produced only small

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.

changes in the predicted probability of the true class (mean $\Delta P(true\ class) \approx +0.02$), confirming that these combinations did not play a critical role in the classifier's final decision.

Across all three pairs, the distribution of $\Delta P(true\ class)$ was centered close to zero for both true invasive and true non-invasive images, indicating a generally balanced impact. The confidence bin analysis (Fig. 4.22b) shows that for true invasive samples, masking slightly decreased the probability of the invasive class at high confidence levels ($0.8 - 1.0$ bin), while low confidence images ($0 - 0.4$ bins) occasionally exhibited small positive shifts. For true non invasive samples, the trend was similarly mild, with small positive shifts at low confidence level and small negative ones at high confidence, implying that the model maintained stable predictions even after trait removal.

These results indicate that *Common* trait pairs are broadly informative features shared across taxa but non decisive cues for distinguishing invasive from non-invasive plants. Their removal causes only minor redistributions of prediction confidence without substantially altering model performance or prediction direction. This limited effect may also reflect the fact that, in many images, not all instances of a trait were removed. Consequently, the model could still rely on remaining occurrences to maintain a stable prediction.

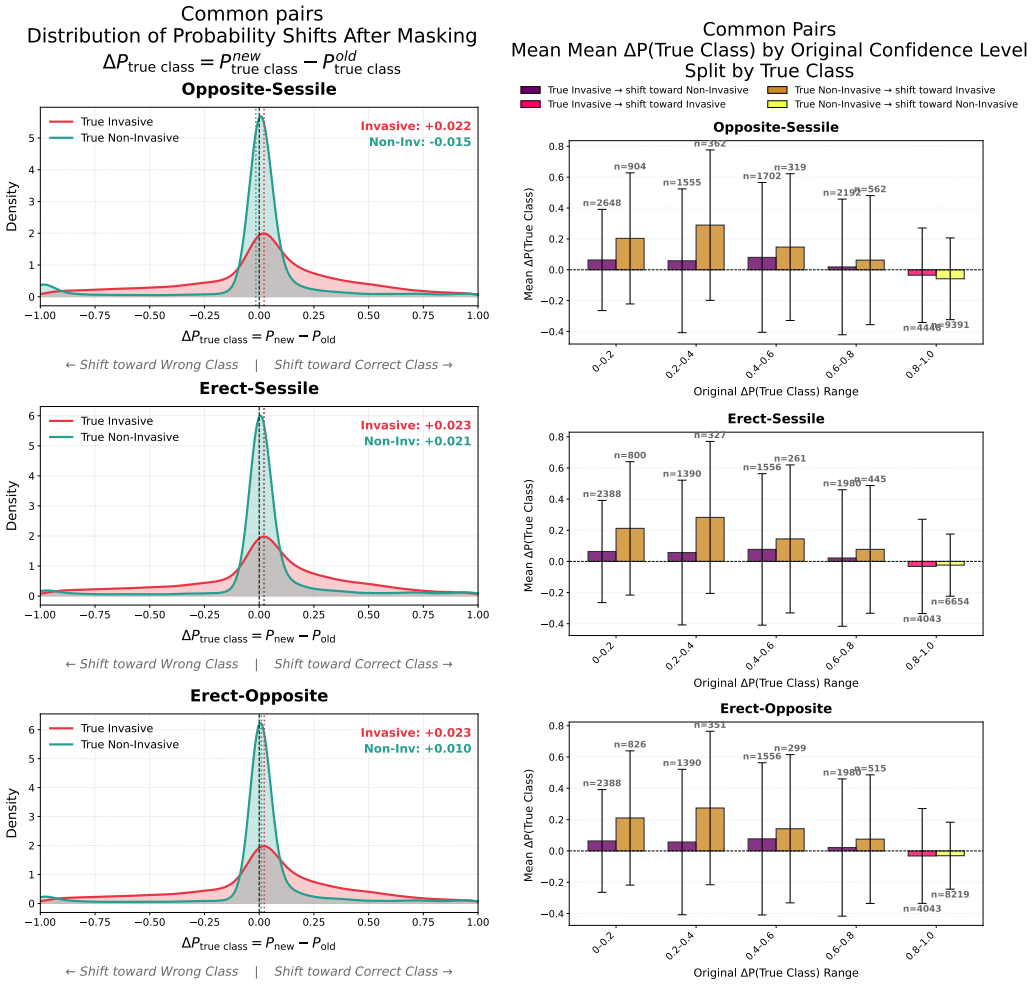
In contrast, pairs belonging to the *Non-Invasive only* category displayed a much stronger and more directional response to masking, as shown in Tab. 4.15 and Fig. 4.23. When *Alternate-Subsessile* regions were masked, the predicted probability for the non invasive class decreased sharply (mean $\Delta P(non\ invasive\ class) = -0.69$), corresponding to a large drop in classification accuracy. As illustrated in Fig. 4.23b, this effect intensified with model confidence: for images the model

originally classified as confidently non-invasive (confidence bins > 0.6), masking consistently caused a large negative shift in ΔP, meaning the model became substantially less confident (or even inverted its prediction) after removal of the traits. This pattern indicates that *Alternate-Subsessile* acts as a strong and distinctive visual cue for non-invasiveness, and its absence drives the model toward invasive predictions.

The second pair, *Linear-Opposite*, exhibits the same general tendency but with smaller magnitude (mean $\Delta P(\textit{non invasive class}) = -0.05$). The per-bin distribution shows a mild positive shift at low confidence and small negative shifts for highly confident images, suggesting a more modest but consistent role in supporting non-invasive predictions. Compared to *Alternate-Subsessile*, this combination appears less diagnostic but still contributes to stabilizing the model's confidence in correctly classified non-invasive samples.

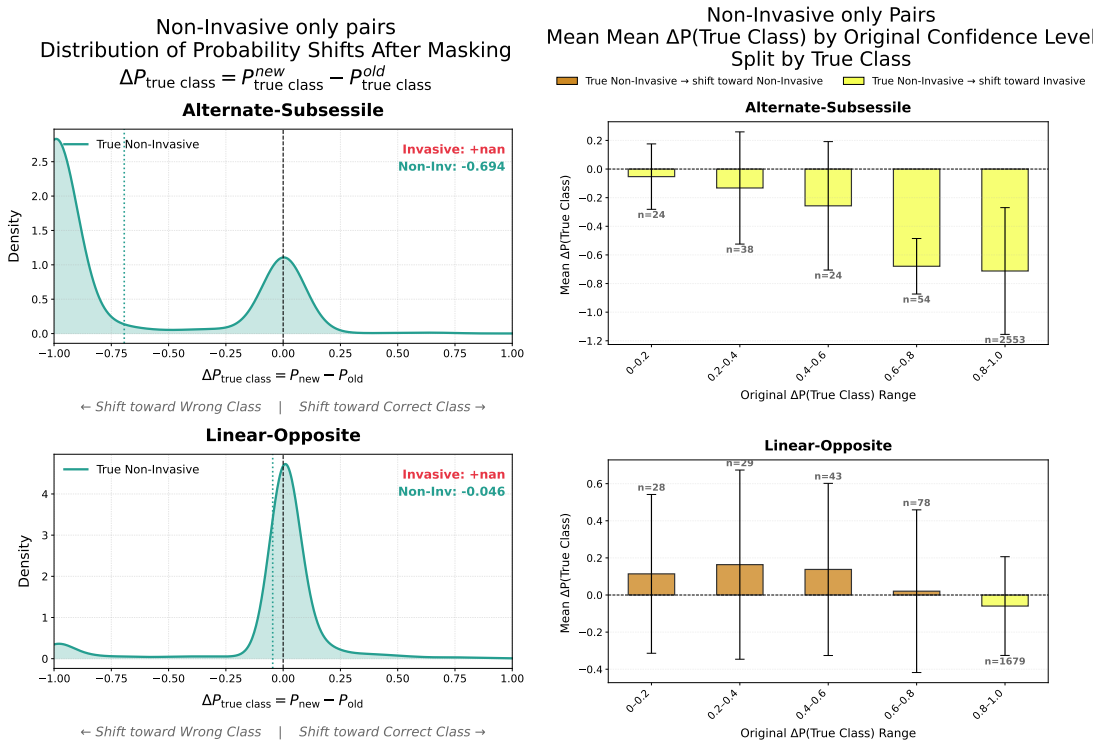
Both *Non-Invasive only* trait pairs produced unidirectional effects on model confidence, confirming that the classifier relied on these combinations as visual trademarks of non-invasive species. Their removal caused a systematic bias toward the invasive class, emphasizing their importance as discriminative, class-specific features.

Overall, these results confirm that the model's predictions are only marginally affected by masking common combinations, which appear to provide redundant information, while the removal of traits distinctive of non-invasive species strongly alters both prediction accuracy and confidence.



(a) **(b)**

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.



(a) **(b)**

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 this thesis, we present a pipeline for identifying morphological traits associated with the potential invasiveness of plant species within a genus, using only image data. We chose the *Lythrum* genus as a case study to determine which visual traits are most informative for distinguishing invasive from non-invasive species.

First, we collected our image data from `iNaturalist.org`, a platform where users can upload pictures of living organisms, making them available for citizen science projects. We then used the state-of-the-art computer vision model BioCLIP-2 to generate embeddings for the images in our dataset. BioCLIP 2 proved effective for this task, distinguishing species with minimal morphological differences and providing fine-grained representations of the data. Using these embeddings, a classifier was trained to predict whether each input image corresponded to an *Invasive* or a *Non-Invasive* species. Specifically, one classifier was trained per species, using all images except those representing the target species, ensuring that the classifier had never seen the species being analyzed.

Next, we incorporated explainability into the pipeline. We applied Integrated Gradients, an XAI algorithm that produces feature attribution maps, to highlight the regions in each image that the model considered most relevant for its prediction. These regions were then clustered and manually annotated, either as one or more biological structures (*Leaf, Flower, Stem*) or as non-informative elements (*Hand, Background/undefined*) depending on what they represented.

With these labeled regions, we extracted patterns and quantified traits associated with invasive or non-invasive species. We aggregated the labeled regions back into the original images and calculated multiple metrics for each. Predictions from the original models were separated by error type (True Positive, True Negative, False Positive or False Negative). A Random Forest classifier was then applied to identify which metrics or image features correlated with prediction accuracy.

Since single features did not yield strong results, we conducted a *pairwise* trait importance analysis. Traits were combined in pairs and grouped according to

class specificity. From the Random Forest, we selected the top five trait pairs by importance and evaluated predictions after masking regions containing those traits. We observed that for three pairs common to both invasive and non-invasive species, accuracy remained largely unchanged. However, for the two pairs exclusive to non-invasive species (*Alternate-Subsessile* and *Linear-Opposite*, all traits describing leaf morphology), masking these regions caused a significant drop in accuracy, with the classifier changing predictions for several images. In particular, masking the traits of the *Alternate-Subsessile* pair led to a 70.5% decrease in accuracy, with 71.3% of the images changing their label from the original prediction to the masked prediction.

This indicates that visual traits can effectively predict non-invasiveness. Due to dataset limitations, we could not identify traits linked specifically to invasiveness, as no exclusive trait pairs existed for invasive species. Nevertheless, the results remain promising and support the validation of our methodology. It is also worth noting that masking any of the top five trait pairs, both common and non-invasive, resulted in a drop in accuracy, highlighting the potential for this approach to be adopted and expanded in future research.

There are several considerations to address regarding the development of this work.

First, we had to remove *Lythrum hyssopifolia* from the analysis because the model was unable to classify it correctly. While this decision improved the overall results, it was motivated by the identification of *L. hyssopifolia* as an analytical outlier rather than a biological one. To the best of our knowledge, there are no studies highlighting anomalies in the invasiveness of *L. hyssopifolia*, nor explaining why the model treats it differently from the other two invasive species in the dataset. Contrary to our initial assumption, this may suggest that either the model struggles to embed *L. hyssopifolia* as a distinct species, or that not all invasive species can be identified solely based on visual traits. We encourage future researchers building on our work to explore this topic further.

The dataset itself also presents potential limitations: after removing *Lythrum hyssopifolia*, only two invasive species remain compared to twenty-seven non-invasive species, even though the sample sizes of the two classes are only slightly unbalanced. This reduces the variability available for the model to learn from, forcing it to rely on just two species and their images. Future studies adopting a similar approach should aim to use a more diverse dataset, which would allow the model to capture a wider range of traits and differences between species, potentially exploring additional genera or families.

Another limitation can be found in the explainability pipeline: the original images were not downloaded at the highest possible quality due to storage constraints and API usage limits. While this does not appear to affect model performance, as shown by our results, it does mean that some of the extracted regions are of lower

resolution, which could impact their embeddings or subsequent analysis.

Manual cluster labeling presents another constraint. Although validated against two algorithms, the labels were based on a subset of regions reviewed by a small group of thesis researchers. This approach was necessary, as no pre-trained segmentation or classification model exists to reliably identify the considered traits from images.

The traits we used to label the clusters were limited and generic, but this was necessary to ensure that we could accurately identify what was included in each region (for example, a flower or a stem). In future studies, it could be valuable to explore the inclusion of sub-traits to enrich the analysis and increase its overall relevance.

Finally, we labeled every species that is invasive anywhere in the world as 'Invasive'. This decision was based on the assumption that interspecific differences are stronger than intraspecific ones (e.g., between native and invasive populations of the same species). Since we had limited geolocalized data, we consciously discarded image location information to access a greater amount of data. Future research could investigate geographic variation, comparing locations where a species is invasive, alien but non-invasive, or native, to determine whether invasiveness is an inherent species trait or context-dependent.

Integrating additional metadata, such as phylogeny, temporal information, or environmental data, could further enrich the analysis. Examining correlations between visual traits and these categorical or numerical variables may reveal new insights into the factors that drive invasiveness.

In conclusion, this thesis demonstrates that it is possible to identify invasive species using only visual traits extracted from images, and highlights the taxon-specific traits that are most relevant for these predictions. This study contributes to the growing body of research on the morphological characteristics of invasive species, supporting their identification and the mitigation of their impacts.

Appendix A

Labels enrichment with species characteristic traits

We enrich the labels inherited from the cluster assignment (see Sec. 4.3.3) using specific traits for each species and for each label. For each region extracted by images of our dataset, we first identified the species it belonged to. Each region is labeled (after the clustering phase) to indicate the presence of key biological structures *Leaf, Flower and Stem*. Using this information, we assigned to the image the characteristic traits associated with the present structures for that species, as defined in Tab. A.1. For example, a region of *Lythrum anatolicum* labeled as containing both flower and 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

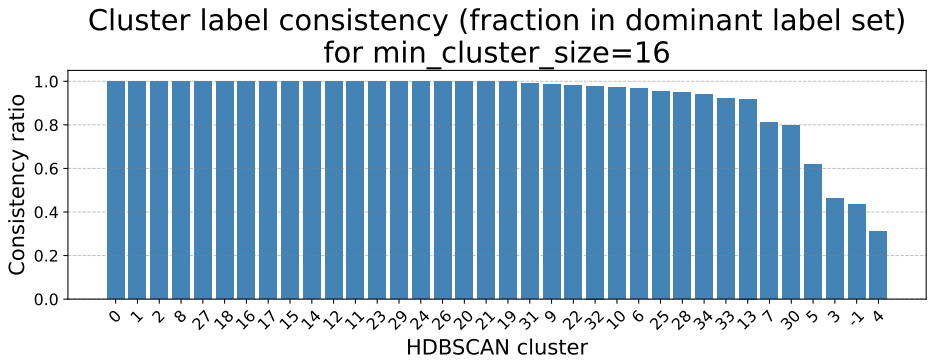
Table A.1: Characteristic traits for each species for each biological structure.

Appendix B

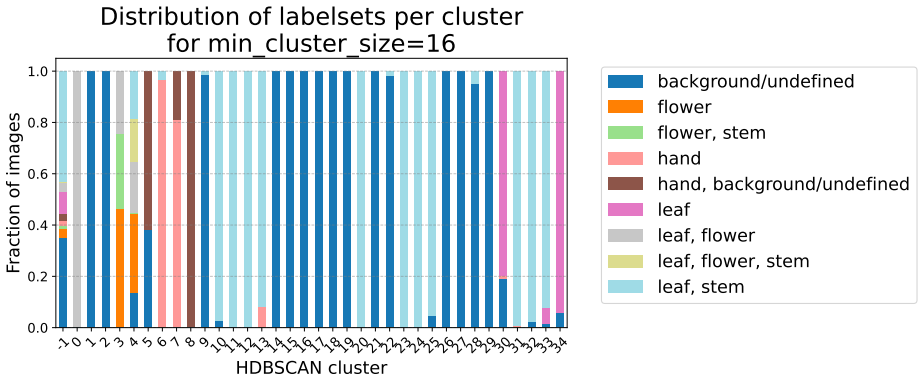
HDBSCAN clustering validation details

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 provide per-cluster analysis for the best (with $min\_cluster\_size = 16$) and the worst (with $min\_cluster\_size = 27$) configurations, respectively. Cluster -1 represents the noise cluster, and is therefore excluded from our analysis. The distribution of consistency ratios (Figs. B.1a and B.2a) align with the gloabl distribution (Fig. 4.16a), as expected. Figs. B.2a and B.2b depict the detailed distribution of labelsets within each cluster. In most cases, even when a cluster contains multiple labelsets, they are closely related. For instance, in Fig. B.2b cluster 2 includes *"flower"*, *"flower, leaf"* and *"flower, stem"*. There are also occasional cases where biological labels are mixed with non-biological ones (for example, cluster 30 in Fig. B.1b has both *"background/undefined"* and *"leaf"*), but usually, one of these represents only a small fraction of the cluster's images. As expected, the $min\_cluster\_size = 27$ configuration yields slightly worse results. This settings produces only 9 clusters (summing it with the noise cluster is just above the threshold): as shown in Fig. B.3, the number of clusters produced by a configuration has a substantial impact on the resulting average consistency ratio. Fewer clusters merge diverse samples, reducing consistency, whereas more clusters capture finer homogeneity.

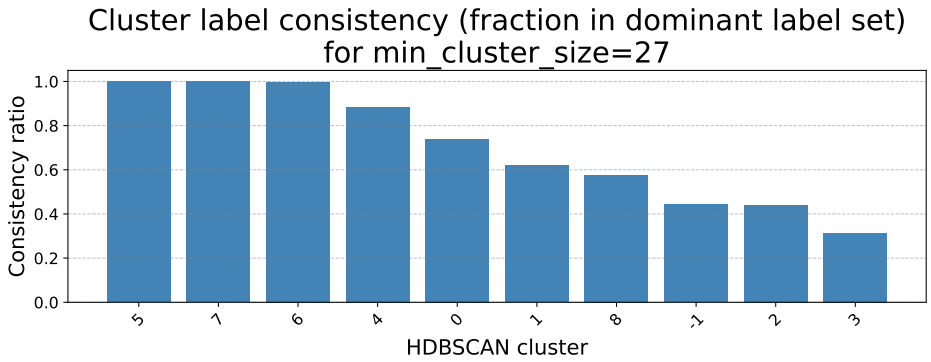


(a)

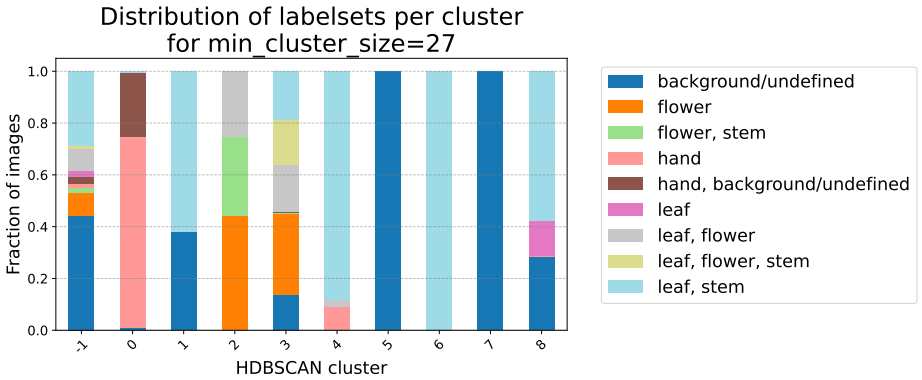


(b)

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.

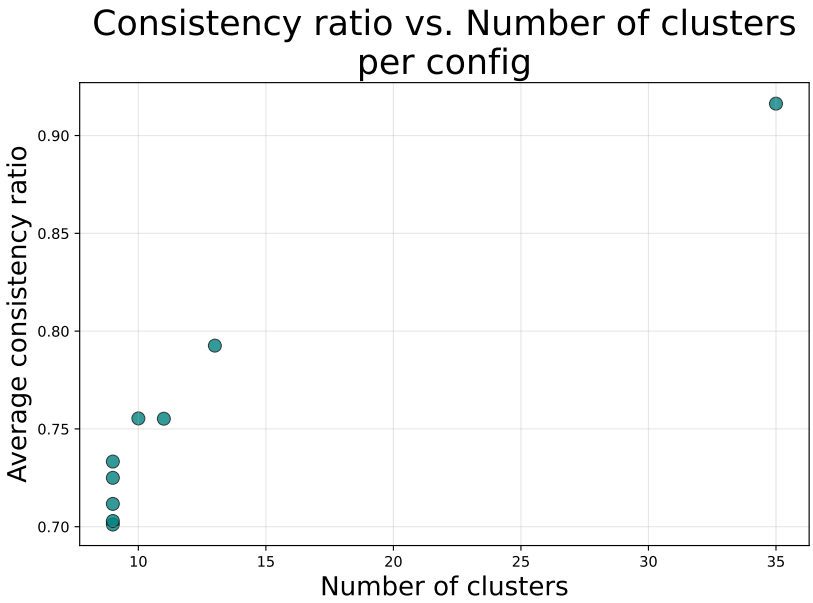


Figure B.3: Correlation between the average consistency ratio of each configuration and the number of clusters generated by the configuration.

Appendix C

Metric-specific correlation with accuracy

This appendix reports the detailed results of the metric-specific accuracy correlation analysis described in Sec. 3.3.1. For each metric, images were divided into discrete categories, and the corresponding mean prediction accuracy, sample size, and Kullback-Leibler (KL) divergence from the overall species distribution were computed. The goal of this analysis was to identify whether particular image level characteristic were systematically associated with classification performance, while also evaluating whether such effects could be attributed to differences in taxonomic composition rather than the metric themselves.

C.1 Pielou Evenness

Accuracy varies across categories of the Pielou evenness index, with a significant ANOVA result ($p < 0.05$), for details see Tab. C.1 and Fig. C.1. Images with *Simple* or *Low* evenness values (i.e., dominated by few traits) show slighlty higher accuracy (up to 0.81) compared to those with *Medium High* or *Very High* evenness, where accuracy drops to around 0.74

However, KL divergence increases notably for intermediate categories, suggesting that these differences may partly reflect changes in species composition.

Overall, the trend may indicate that images with more homogeneous trait distributions are classified more reliably, although this effect is 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, mean accuracy for the category, sample size and KL divergence from the dataset distribution is shown.

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 ($p < 0.05$).

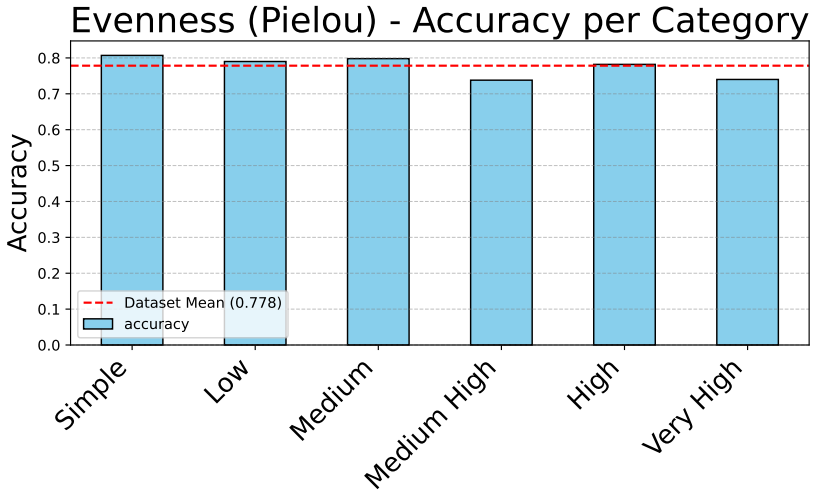
Category	Accuracy	Sample size	KL Divergence
Simple	0.790	2574	0.041
Low	0.885	3963	0.620
Medium	0.762	10390	0.072
Medium High	0.854	3954	0.580
High	0.764	11057	0.141
Very High	0.609	2451	0.759

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, mean accuracy for the category, sample size and KL divergence from the dataset distribution is shown.

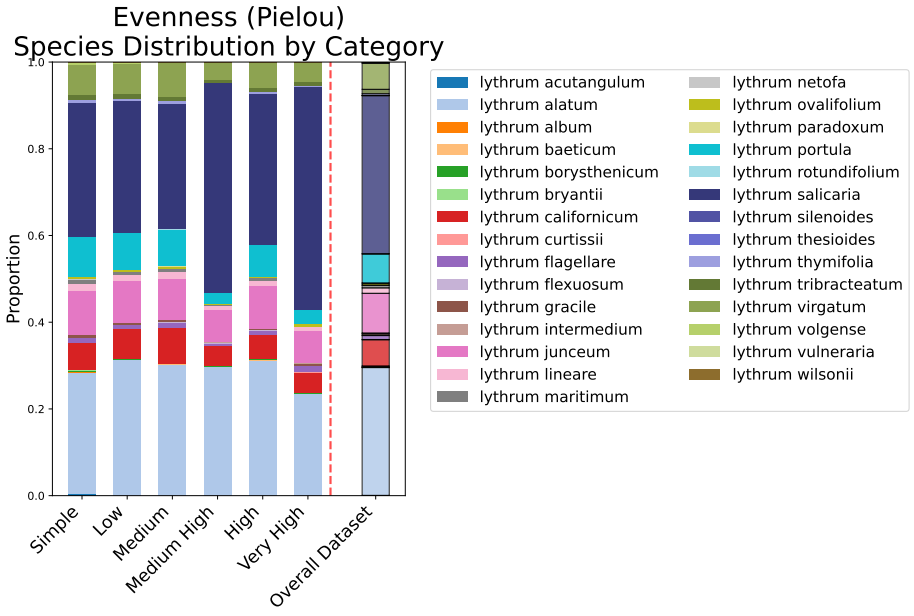
ANOVA: $F = 170.03$, $p \approx 0.00$; the highest mean accuracy is observed in the Low category (0.885) and the lowest in the Very High category (0.609). Accuracy differences are statistically significant ($p < 0.05$).

C.2 Distinct Traits (richness)

Richness shows a statistically significant association with accuracy ($p < 0.05$), but the pattern is irregular and influenced by species imbalance (see Tab. C.2 and Fig. C.2). *Low* and *Medium High* richness categories reach the highest accuracies (around 0.85-0.88), while both *Simple* and *Very High* richness values correspond to reduced accuracy (0.61-0.76).



(a)

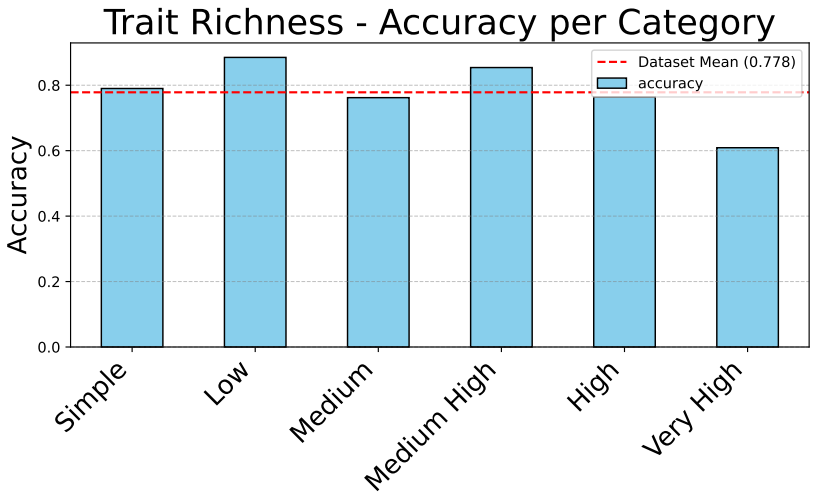


(b)

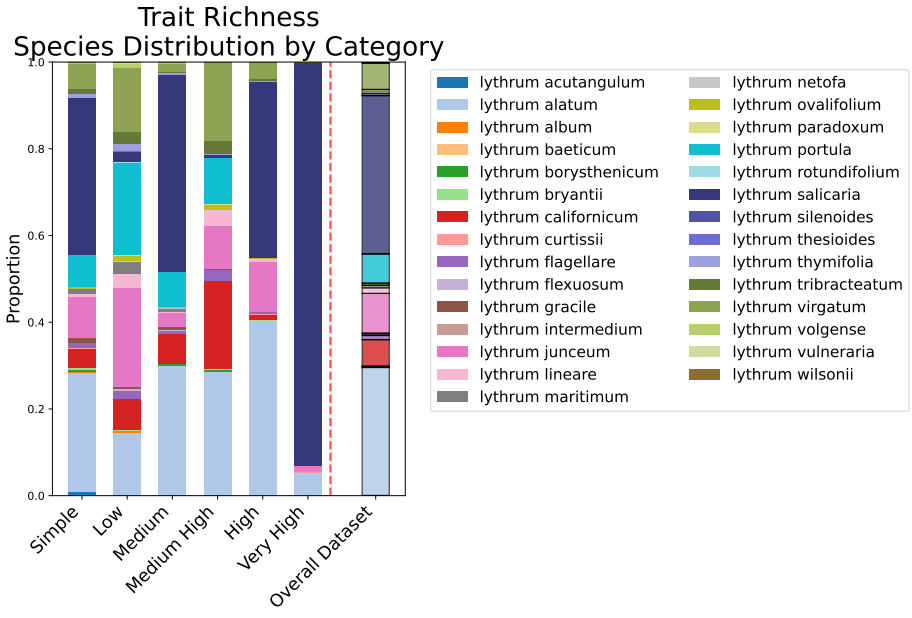
Figure C.1: For the Pielou Evenness analysis, accuracy for each category (a) and distribution of species for each category (b) is shown.

However, the corresponding KL divergences are large (up to 0.76), indicating that bins with extreme richness values are dominated by specific taxa.

Consequently, while images with moderate trait diversity appear more stable for classification, this trend should not be interpret as causal.



(a)



(b)

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

C.3 Hand Fraction

Accuracy remains remarkably stable across all categories of hand fraction, ranging from 0.77 to 0.78, with no statistically significant differences (see Tab. C.3

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, mean accuracy for the category, sample size and KL divergence from the dataset distribution is shown.

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.

and Fig. C.3).KL divergence values are consistently low.

These results indicate that the presence of human hands within the image does not systematically bias the model's predictions.

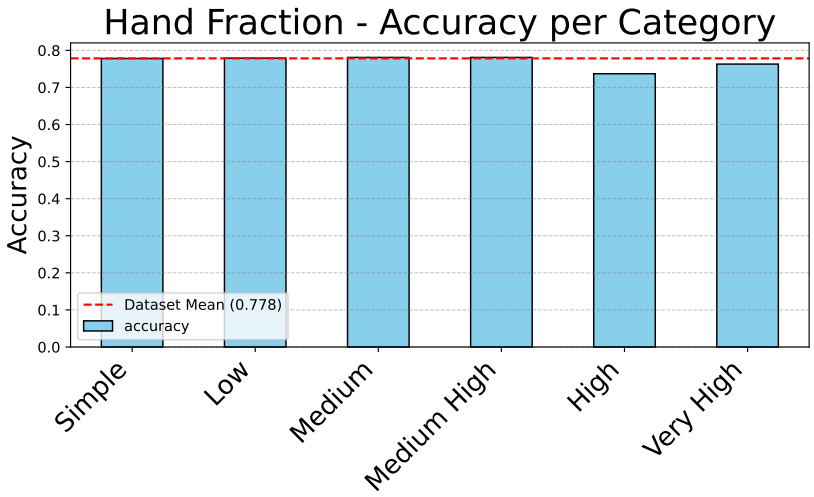
C.4 Background/undefined fraction

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

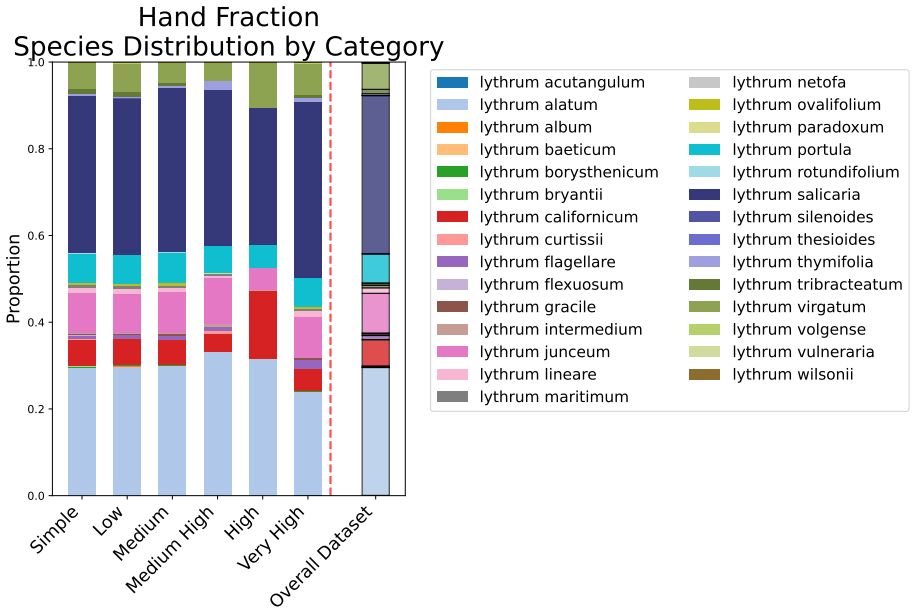
Table C.4: Results of Background/Undefined fraction correlation with accuracy analysis. For each region coverage category, mean accuracy for the category, sample size and KL divergence from the dataset distribution is shown.

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.

Similarly, Tab. C.4 and Fig. C.4 display that the fraction of regions labeled as background or undefined shows no consistent relationship with classification



(a)

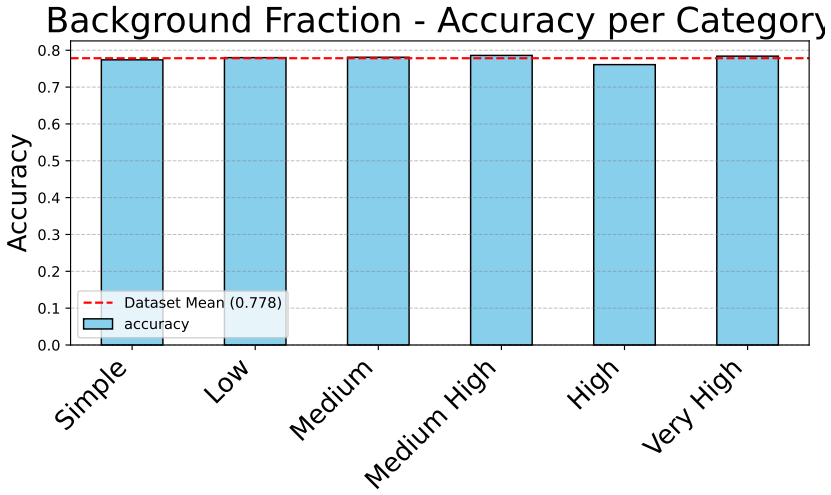


(b)

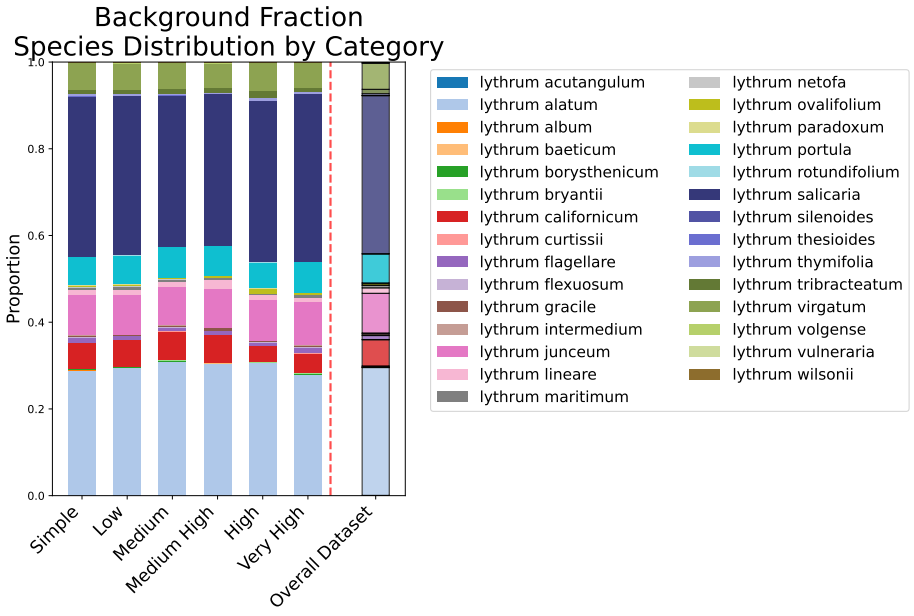
Figure C.3: For the Hand fraction analysis, accuracy for each category (a) and distribution of species for each category (b) is shown.

accuracy. Accuracy values fluctuates minimally (0.76-0.79) across categories, and KL divergence remains below 0.3.

This confirms that variation in background area does not influence model correctness, suggesting that the model effectively focuses on relevant plant regions.



(a)



(b)

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

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

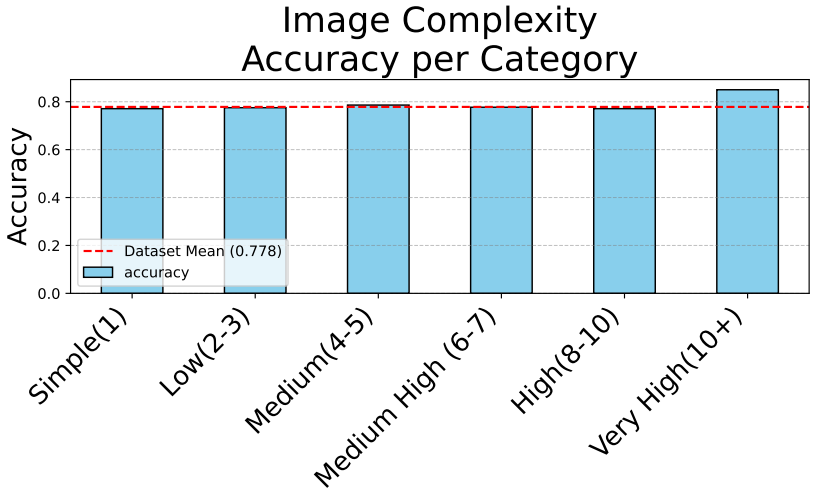
Category	**Accuracy**	**Sample size**	**KL Divergence**
Simple (1)	0.771	3978	0.008
Low (2-3)	0.775	12968	0.002
Medium (4-5)	0.786	10680	0.001
Medium High (6-7)	0.777	4820	0.007
High (8-10)	0.771	1729	0.026
Very High (10+)	0.850	214	0.098

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

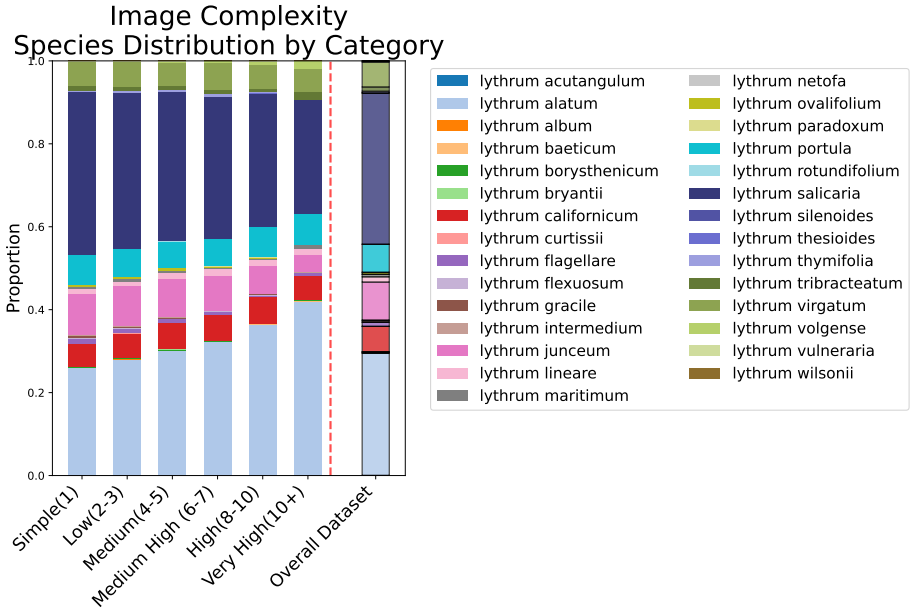
ANOVA: $F = 2.554$, $p = 0.026$; the highest mean accuracy is observed in the Very High (10+) category (0.850) and the lowest in the Simple (1) category (0.771). Accuracy differences are statistically significant ($p < 0.05$).

tends to increase slighlty with complexity, reaching its highest value (0.85) for very 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 for the model, the observed effect is limited.



(a)



(b)

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

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, 12).

[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, 12).

[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, 12).

[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 Jiří 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. 6).

[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. 6).

[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. 6).

[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, 12, 13, 20).

[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, 11, 13, 16, 20).

[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. 7).

[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. 7).

[27] Imageomics institute. *Imageomics*. [Online; accessed 8-September-2025]. 2025. URL: `https://imageomics.osu.edu/about` (cit. on p. 7).

[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. 8).

[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. 8).

[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. 8, 14).

[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, 14).

[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. 12).

[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. 12).

[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. 12).

[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 pp. 12, 13).

[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. 16).

[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. 19).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 26).

[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. 32).