Automated Analysis of the Vienna "Naturhistorisches Museum" Herbarium Collection

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Mr. Marko Kadić, BSc
Matrikelnummer 12045128

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Senior Lecturer Dipl.-Ing. Dr.techn. Michael Reiter
Mitwirkung: Dipl.-Ing. Dr.techn. Florian Kleber

Wien, 1. Dezember 2024

Marko Kadić

Michael Reiter

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Automated Analysis of the Vienna "Naturhistorisches Museum" Herbarium Collection

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Mr. Marko Kadić, BSc
Registration Number 12045128

to the Faculty of Informatics

at the TU Wien

Advisor: Senior Lecturer Dipl.-Ing. Dr.techn. Michael Reiter
Assistance: Dipl.-Ing. Dr.techn. Florian Kleber

Vienna, 1st December, 2024

Marko Kadić

Michael Reiter

Erklärung zur Verfassung der Arbeit

Mr. Marko Kadić, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. Dezember 2024

Marko Kadić

Acknowledgements

I would like to express my sincere appreciation to my supervisors, Dr. Michael Reiter and Dr. Florian Kleber, for their invaluable guidance, support, and constructive feedback throughout this research process. Their expertise and encouragement were essential in shaping this work.

I am also thankful to my family and friends for their unwavering support and patience during this journey. Their encouragement kept me motivated during challenging times.

Finally, I would like to acknowledge The Vienna University of Technology and Vienna Scientific Cluster for providing the resources and support necessary to complete this thesis.

Abstract

A herbarium is a collection of preserved plant specimens with information from the collector and additional details on the plants. The Herbarium of the Natural History Museum in Vienna (NHMW) was established in 1885 and is now ranked amongst the top ten botanical collections in the world. Current holdings in the NHMW are approximately 5.5 million plant specimens. The herbarium is especially rich in types (the plant specimen to which the scientific name of that species is attached, to serve as reference for all other specimens that belong to the species) with around 200,000 type specimens.

The digitized images can easily be provided to botanical researchers. However, efficient search tools are needed to explore the collection of 5.5 million specimens. Thus, the handwriting of the collector on the herbarium sheet itself, the provided metadata, the layout of the label, and the dried plants on the sheets themselves can be used as information. The handwriting holds the information on the specimen created manually by the botanical researchers, such as the family, genus, species classification, and in some cases the year, location and the person who made the herbarium specimen entry.

The overall goal of this diploma thesis is to develop deep learning-based methods that can model the similarities of plants based on the herbarium specimen images (including besides the specimen image all additional information of the image, such as the handwritten or printed label information created by the botanical researchers and the metadata provided by the NHMW).

The deep learning technique that is utilized is *contrastive learning*, a method for self-supervised learning that does not require labeled data to cluster similar data points together. Using this method instance discrimination (clustering similar data points together) is done on the unlabelled data of the specimen images, after which the plant specimens are classified into their respective families, genera and species.

Herbarium dataset ground truth data is usually entirely private to the collection (noted exception being the Herbarium Challenge 2022 dataset), making it hard to apply supervised models in a general setting, as different herbaria can consist of entirely different families, genera and species. Contrastive learning is used in this thesis to mitigate this problem as well as to develop a more robust and general system of classification. Secondly, it has achieved state-of-the-art results on the ImageNet dataset, and thus has also been chosen as the method for this thesis in order to evaluate how well the technique performs

on the downstream task of plant species classification compared to other (fully supervised) approaches already used to tackle this problem.

A visualization application is created within this thesis for botanical research analysis which uses the developed deep learning based classification. The results of the classification and the ground truth information of corresponding images are made available to the researchers through the interactive visualization tool.

The ground truth data has been made available by the NHMW for a dataset of 1078 specimen images, along with additional information about these specimens (location, time and name of the collectors). In addition to the NHMW dataset, the Herbarium 2022 dataset is also used as a benchmark for the quantitative analysis of the contrastive learning method, the dataset consists of 1.05 million images with the ground truth information on the plant species, genus and family.

In addition, the recognition of the printed and the handwritten label content is evaluated based on state-of-the-art Handwritten Text Recognition (HTR) and Optical Character Recognition (OCR) solutions, and evaluated on the NHMW herbarium dataset. This is done to show the applicability of text recognition to this type of herbarium specimen analysis. Error rates are calculated by comparing the recognized (predicted) text from the herbarium specimen labels in the plant specimen images with the ground truth data provided by the NHMW.

The main contribution is the use of a self-supervised, deep learning, methodology in training a model for herbarium plant specimen classification and its comparison to supervised approaches, and to the state-of-the.art approaches. The main advantage of a self-supervised methodology is that no target class labels are needed in the embedding model training while still utilizing the advantages of convolutional neural networks over the feature engineering methods used in a self-supervised and unsupervised setting. Further contributions of the thesis include a web application that allows the researchers to look into the clusters of specimens that belong together based on the contrastive learning model embeddings, and select individual images of specimens to further look into (paired with the already available data on the specimens that the NHMW holds), a newly created dataset for the segmentation of the dried pressed plant parts of the specimen images and an evaluation of HTR and OCR methods used for information extraction form the specimens.

Contents

CHAPTER 1

Introduction

Herbaria contain a systematic overview of plant specimens collected, preserved, and well-documented by botanists for future use. Herbaria are the foundation of botanical research, and they are used for a variety of studies, like plant taxonomy, species variability, extinction risk, and phenological trends[4][3][46]. These physical specimens ensure reproducibility and unambiguous referencing of research results relating to organisms.

An excerpt image from the NHMW herbarium can be seen in Figure 1.1. There are five main parts to a herbarium specimen:

- The dried and pressed plant that is attached to the sheet
- Plant label (printed or handwritten)
- The color reference chart exists to ensure image quality in the process of digitization of the specimens
- The scale bar provides the reference information of the actual physical measurements of the specimen
- The white strips are there to hold the specimen attached to the paper

Recent initiatives, Natural History Museum Vienna (NHMW) included, started to digitize this information and make it available to botanists and the public through web portals and web services. When identifying a plant's taxonomy[42], the researchers determine its family, genus and species, following the standard Latin binomial system of nomenclature. The first name represents the genus and is capitalized, the second name identifies the species and is written in lowercase, and the entire name is italicized. The NHMW herbarium holds a collection of 5.5 million specimen. The herbarium is especially rich in *type specimens*. The information about the plant species that is annotated by the

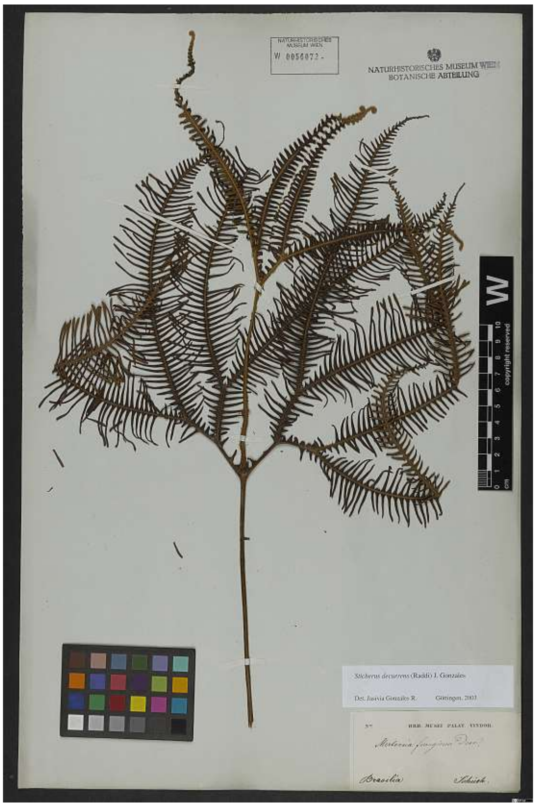


Figure 1.1: Showing a herbarium specimen

researchers is dependent on their knowledge of the species, where the type specimen serves as a prototype of that species class. The "type specimen" is the most accurate specimen of a plant species, thus being the center reference for the classification of all other plant specimen that belong to a species. Therefore the task is a form of instance discrimination in itself for the botanical researchers, making "type specimen" an important part of every herbarium.

According to Rojas et al.[5] *"Thousands and thousands of sheets are still not identified at the species level while numerous sheets should be reviewed and updated following more recent taxonomic knowledge. These annotations and revisions require such a large amount of work from botanists that it would be unfeasible to carry them out in a reasonable time."* Computer vision approaches based on the automated analyses of these sheets are therefore necessary for such species identification tasks. These tools should also strongly accelerate evolutionary and ecological studies by providing quick access to the most interesting specimens of a family/genus/species group of interest. A visualisation application that, based on herbarium sheet images across multiple collections worldwide, finds the plant specimens more similar to a candidate will be of great help for the researchers at the NHMW to solve specific research questions as pointed out before 6.

This thesis focuses around the following research questions:

- RQ1: What is a suitable deep learning architecture for the analysis of the herbarium plant specimens?
- RQ2: How can state-of-the-art text recognition (handwritten and printed) methods be used for herbarium data analysis?
- RQ3: How well do self-supervised contrastive learning methods perform on the task of plant species classification?

1.1 Deep Learning on plant images

The first problem that is addressed is the automatic analysis of the dried and pressed plant region of the images. The variety of different plants and plant features makes the task of categorizing and systemically ordering the plants difficult and time-consuming[20], which is why researchers tried to apply machine learning and deep learning methods to the herbarium collections[29]. Some of the studies focus on the classification of the plants within their genus and species[5][45][17]. Botanical researchers would also like to use the data to gain deeper insights into the development and the phenology of the plants such as: flowering, leaf unfolding (or budburst), seed set and dispersal, and leaf fall in relation to climatic conditions[46]. Meaning that, unlike in classification, the model has to learn robust representations of individual plants, and not only the shape of the plant specimen. Therefore the specimen image analysis focuses on the classification of the respective plant species, with the goal of applying the learned representations on the downstream task of extraction of the phenology of the plant specimens. contrasive learning is a self-supervised approach to deep learning that does not require labeled data to cluster similar data points together, while still using the advantages of Convolutional Neural Networks (CNNs), and is used to tackle the aforementioned problem. Herbarium dataset ground truth data is usually entirely private to the collection (noted exception being the Herbarium 2022[28] dataset), making it hard to apply supervised models in a general setting, as different herbaria can consist of entirely different genus types. Contrastive learning is used in this thesis to mitigate this problem as well as to try and create a more robust and general system of classification. Secondly, it has achieved state-of-the-art results on the ImageNet dataset[18][8] and thus has also been chosen as the method for this thesis in order to evaluate how well the technique performs on the downstream task of plant species classification compared to other (fully supervised) approaches already used to tackle this problem[43][29].

1.2 Label character recognition

The plant label holds the annotation information about the specimen, annotated by the botanical researchers that collected the specimen. The information it holds can be split into two categories. Firstly, the genus name and the species name of the plant are present in 1078 samples, of the 1103 images provided by the NHMW. The second set of annotation information is the place where the sample was taken (geolocation), the name of the person who made the sample, and the date when the sample was taken, and it is present only in some samples (different amount for each of the information categories). The details about the dataset are in the datasets section of the results section 4.1.1 of the thesis. The label itself can be printed or handwritten.

Therefore a tool to retrieve the textual information of the labels is useful to the researchers in retrieving the data about already made annotations, that have not yet been transcribed.

The plant genus and plant species name information (retrieved from the handwritten or printed labels) is useful to the botanical researchers. They can use the transcription model alongside a model that only looks at plant phenology and clusters similar plants closer together, to discriminate plants into species and genera automatically.

Since the classification of the plants carried out by the botanists is dependent on *type specimens* and the boundaries between species are often unclear, combining the knowledge of botanists (human annotations of the plant species) with deep learning models - see 3.1 can result in drawing clear boundaries between species in their classification, or a change in an existing boundary. The comparison can also provide further insights into the differences/similarities between similar plants of the same species (classified by the researchers manually in the herbarium text labels).

The thesis only provides an evaluation of state-of-the-art HTR and OCR methods to retrieve the text, as a pointer to the future work of HTR in herbarium specimens. The recognition of the printed and the handwritten label content is evaluated based on state-of-the-art HTR solution, and evaluated on the NHMW herbarium dataset. This is done to show the applicability of text recognition to this type of herbarium specimen analysis. Error rates are calculated by comparing the recognized (predicted) text from the herbarium specimen labels in the plant specimen images with the ground truth data provided by the NHMW. An HTR solution called Google Vision HTR[23] is used as well, to show the results that out-of-the-box solutions achieve on the task. The same solution that was used for the HTR is used for the OCR of the Printed Information. The printed information is scraped using an OCR (Optical Character Recognition) solution as well, and the results are compared to the HTR solution. The engine that is used is the Tesseract[35]. This engine is used because it is a standard OCR solution.

CHAPTER 2

Literature research

This section provides an overview of the current state-of-the-art methods in the area of machine learning/deep learning methods for plant family/genus/species classification, deep learning methods for phenological trait detection/extraction of plants, and finally the state-of-the-art methods for handwritten text recognition.

2.1 Deep Learning methods for plant classification

Plant classification of the plants done by the botanists entails the categorisation of each plant specimen into a correct family, genus and species that the plant belongs to (in that hierarchical order). Each plant belongs to a family, a genus within that family and a species within that genus. The classification is time-consuming[20] and dependent on *type specimens* and the boundaries between species are often unclear.

Rojas et al.[5] reported 79.6% accuracy in the classification of the plant family, genus, and species using CNNs, on their own datasets of herbaria properly curated for machine learning purposes, including one small dataset (255 species, 7,500 images) and one large dataset (1,204 species, 260,000 images). An example of a herbarium specimen can be seen in Figure 1.1, only the plant image of the specimen is used in all the papers mentioned. This approach beat the previous state-of-the-art approaches that used feature extraction and classical machine learning techniques. Walker et al.[45] suggested that triplet networks produce good representations, outperforming the CNN models on the task of discriminating between *Syzygium* and *Dendrobium* (distinct genera): 99.9% triplets to 95.3% CNNs accuracy score and *Syzygium* and *Eugenia* (similar genera): 99.9% for triplets and 87.9% for CNNs, Walker used ResNet-18 as the base network, the images were taken from the Herbarium 2021 dataset. Chulif and Chang's work[10] describes a Convolutional Siamese network to learn feature similarities between herbarium images and field plants. They report a value of 0.121 and 0.11 mean reciprocal rank (MRR), on the Herbarium 2021 and Herbarium 2022 datasets respectively, for cross-domain species identification. The current state-of-the-art approaches for genus classification use vision transformers[19]. The "Conviformers: Convolutionally guided Vision Transformers" paper[43] from 2022 used Vision Transformers to approach the task of plant genus classification. The paper achieved the current state-of-the-art results on the Herbarium 202x and iNaturalist 2019 datasets, with an accuracy score of 0.729 on the Herbarium 2021 dataset (2.5M images, 65,000 species) and a 0.824 accuracy score on the Herbarium 2022 dataset[43] (1.2M images, 15501 species).

2.2 Machine learning methods for phenological trait detection and extraction

Phenological traits of plants are characteristics that relate to the timing of events in a plant's life cycle, such as flowering, seed dispersal, and seed maturation[34]. These traits are monitored by looking at plant organs during time and are important for understanding how plants interact with their environment and other organisms. Phenological traits detection and extraction include leaf measurements[14], morphological traits of the specimens[49], detection of plant organs[47], and measurements of reproductive organs[34].

Some of these studies have proposed image processing algorithms[13], while others have used deep learning approaches due to the complexity of target information[34]. The earlier attempts at solving this problem have used various image processing techniques like SIFT descriptors, HoG (Histogram of Oriented Gradients) descriptors, contour models, etc. to automatically locate leaf teeth and extract features like leaf tooth perimeter, area, internal angles, count and other derived features[13]. Despite the successful extraction of the features the proposed algorithms were deemed limited in their capabilities and did not generalize well over different plant taxa. This limitation has been partially mitigated in the new approaches by the use of deep learning techniques that can learn important filters and can generalize well across different plants. This work is only focused on the deep learning approach. These methods have been limited by the fact that these deep learning methods use supervised learning and are dependent on the ground truth data that is usually entirely private to the collection (noted exception being the Herbarium 2022[28] dataset), making it hard to apply these models in a general setting, as different herbaria can consist of entirely different genus types. Contrastive learning as a self-supervised method is used in this thesis to mitigate the problem of using ground truths as well as try to implicitly create a more robust and general system of similar trait detection. Contrastive learning does instance discrimination and clusters the most similar plants together, which implicitly creates clusters of similar phenological traits, meaning that while the same species might appear different in different life cycle in the dataset they can be clustered closer together with another plant species that looks similar to it in a certain life cycle (based on their phenological traits).

2.3 Text Recognition (HTR and OCR)

The current state-of-the-art solution for the HTR problem is DAN: a Segmentation-free Document Attention Network[11]. DAN has been trained on the READ2016[40] dataset, which is composed of early modern German texts. The solution achieved competitive results on the READ 2016 dataset at the page level, as well as double-page level with a CER (Character Error Rate) of 3.43% and 3.70%, respectively, the WER (Word Error Rate) on the READ 2016 dataset was 13.5% (page level) and 14.15% (double-page). DAN has the top results in all categories on the READ 2016 dataset, except for the WER on a line level where it gets outperformed by the Vertical Attention Network[12]. From out-of-the-box HTR solutions (Google, AWS, Microsoft, Tesseract and SimpleHTR), according to benchmark tests[2], and tested on the NIST Hand printed Forms and Characters Database[25], Google Cloud Vision achieved the best score of 9.00 average CER and 90.44 Average Match Score. Additionally, Tesseract-OCR engine[39], achieved a 5.4% CER on the UNLV Fourth Annual Test of OCR Accuracy dataset[37].

CHAPTER 3

Methodology

This section presents the methods and approaches used in the thesis in order to answer the research questions. It is split into two main parts: contrastive learning on the plant images and text detection and extraction. The contrastive learning part goes over the SimCLR method and the methods used subsequently to do classification and segmentation. Herbarium dataset ground truth data is usually entirely private to the collection (noted exception being the Herbarium 2022[28] dataset), making it hard to apply supervised models in a general setting, as different herbarium collections can consist of entirely different plant families, genera and species. Contrastive learning is used in this thesis to mitigate this problem as well as to try and create a more robust and general system of classification. Secondly, it has achieved state-of-the-art results on the ImageNet dataset[18][8], and thus has also been chosen as the method for this thesis in order to evaluate how well the technique performs on the downstream task of plant species classification compared to other (fully supervised) approaches already used to tackle this problem[43][29].

3.1 Contrastive Learning

Contrastive learning, a self-supervised method for representation learning, is used for the task of clustering the plants into semantically rich categories based on their visual features. The goal is a semantically rich general representation of herbarium specimens suitable for downstream tasks like classification or detection of plant features like specific plant organs etc. These representations are needed in order to support the botanical research with a tool that can provide further insights into the development of plants, either through time or through their geographical location (See 4.4). The categories to look for are the reproductive and other organs of plants including, but not limited to: leaves, blossoms, and fruits.

3.1.1 Introduction

Contrastive learning is a self-supervised, task-independent deep learning technique that allows a model to learn data representations without labels. The model learns to differentiate between similar and dissimilar data points by contrasting them against each other. A positive pair in contrastive learning typically refers to two data points that are similar or related in some way, e.g. two different augmentations of the same image, or an image and its textual description (e.g. in CLIP[36]). A negative pair typically consists of two data points that are different or unrelated, e.g. two different images. These augmentations require an anchor, an anchor is a reference data point used to compare other data points against. It serves as the basis for forming positive and negative pairs. The loss function in contrastive learning uses the anchor to compute distances between the anchor and the augmented version (in a positive pair) or the anchor and the example from a different class (in a negative pair). The goal of the contrastive learning is then to minimize the distance for positive pairs while maximizing the distance within negative pairs. The variants of the loss function include: contrastive loss, triplet loss (used in supervised settings), and Normalized Temperature-scaled Cross Entropy (NT-Xent)[7] etc. (as detailed in Section 3.1.2).

The specific implementation of the contrastive learning chosen for this topic is the SimCLR (Simple Contrastive Learning of Representations), a contrastive learning method for unsupervised representation learning proposed by Chen et al. in 2020[7], the paper shows that SimCLR can achieve state-of-the-art performance on image classification benchmarks such as ImageNet and CIFAR-10, as well as on other computer vision tasks such as object detection and segmentation.

In this thesis the representations are learned from unsupervised SimCLR training. These representations are assessed using two ways of linear probing: Firstly, logistic regression[16] and secondly, a K-Nearest Neighbors (KNN) classification[21]. Additionally, after SimCLR training the base network was frozen and fully supervised fine-tuning was performed on the CNN model to evaluate the updated network. These methods are discussed further in the SimCLR implementation details section of the thesis. The method is one of the current state-of-the-art approaches for pure contrastive learning[33], the latest being[41] that adds another loss function on top of the SimCLR approach. SimCLR has recently been beaten by Vision Transformers[48] and by a combination of Vision Transformers and Contrastive Learning[9] on the ImageNet dataset.

Other notable variants of contrastive learning include supervised contrastive learning (SupCon) that uses labels and then uses augmented pairs to push positive anchors closer together by the augmentation features, and CLIP (Contrastive Language-Image Pretraining)[36] that learns visual representations from language labels by contrasting image-text pairs. It is trained on a dataset of images with corresponding textual descriptions, pairing images with their relevant textual descriptions to enable transfer to various visual tasks without the need for specific training on specific datasets.

Learning image features without labels, using deep learning methods, is a novel idea and the task of plant species classification provides an opportunity to research the performance of contrastive learning and compare its performance to other deep learning methods, already used on this task. Another task of the deep learning approach is to explore the similarities between plants in terms of their phenology and explore the clusters, and not only approach the classification task. The SimCLR method is used, still ranked top 5 on ImageNet top 5 accuracy[8] to acquire these general representations of herbarium specimens to use them in downstream tasks, and in botanical research as well.

SimCLR uses an image, called an anchor, to create two augmented versions of the image (a positive pair), these positive pairs are compared to all other image augmented pairs (negative pairs) and a loss function calculates the similarity between positive and negative pairs, making the minimization of the said function the goal of the training. The method uses a siamese network architecture to learn representations by contrasting similar and dissimilar pairs of augmented views of input data. This means that each image of the pair is passed through the same network before the distance between the pair is minimized, and the weights are updated, making the network, conceptually, a siamese twin 3.1. The method is described in more detail in the next section.

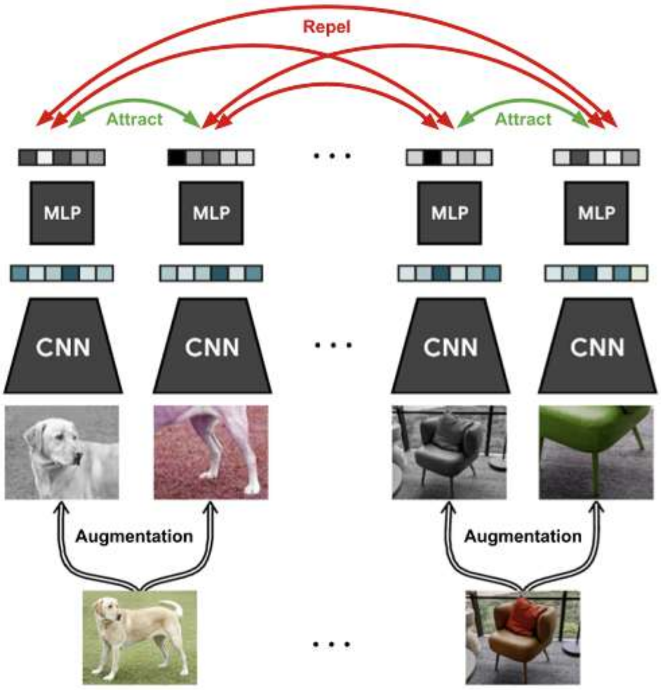


Figure 3.1: Contrastive Learning Concept

The advantages of SimCLR are learning representations from and the ability to learn more robust and generalizable representations by leveraging the diversity of unlabeled data and using multiple augmentations of the same image. These representations obtained from the contrastive learning can be used for downstream tasks such as, in our case, herbarium specimen plant species classification.

In conclusion, the SimCLR contrastive learning method is an approach for unsupervised representation learning that can improve the performance of deep learning models on the task of the automated analysis of the herbarium specimens.

3.1.2 SimCLR

At each iteration, we get for every image x two differently augmented versions of the original image, We refer to the augmentations as $\tilde{x}_i$ and $\tilde{x}_j$, the augmentations of the image that have been shown to work best (see[7]) are a random crop of the image along with the color augmentations of each image crop. The augmentations differ by experiment, but all include random resize and crop of the original image, for more details refer to the results section of the thesis.

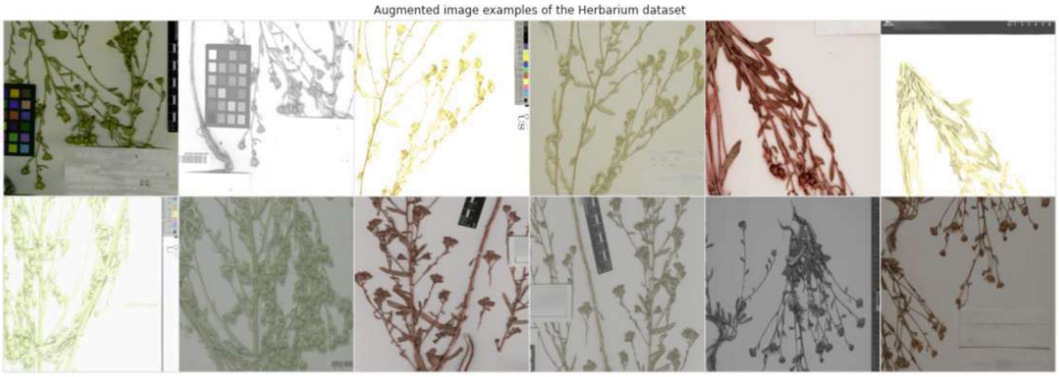


Figure 3.2: Example Augmentations of Herbarium 2022 dataset

The example of cropped pairs can be seen in Figure 3.2. Both of these images are encoded into a one-dimensional feature vector, $\tilde{h}_i$ and $\tilde{h}_j$ respectively, by passing them through a CNN network, referred to as $f(\cdot)$, the vectors are of 512 dimensions in the case of encoding the image with ResNet18 and 1024 in the case of ResNet50. We want to maximize similarity between these two augmentations of the image, as they both belong to the same class (original image), minimizing it to all other images in the batch.

The encoder network is split into two parts: a base encoder network $f(\cdot)$ (ResNet18, ResNet50), and a projection head $g(\cdot)$. The base network is responsible for extracting a representation vector from the augmented data examples. ResNet18 and ResNet50 architectures were used as $f(\cdot)$, and we refer to the output as $f(\tilde{x}_i) = h_i$.

The projection head $g(\cdot)$ maps the representation h into a latent multidimensional vector space where we apply the contrastive loss, i.e., compare similarities between vectors.

Algorithm 3.1: SimCLR Main Learning Algorithm

```
Input  : Batch size $N$, temperature $\tau$, encoder $f$, projection head $g$, set of
         augmentations $T$
Output : Trained encoder network $f(\cdot)$
foreach sampled minibatch $\{x_k\}_{k=1}^{N}$ do
    foreach $k \in \{1, \dots, N\}$ do
        draw two augmentation functions $t \sim T$, $t' \sim T$;
        // first augmentation
        $x_{2\tilde{k}-1_k} = t(x_k)$;
        $h_{2k-1} = f(x_{2\tilde{k}-1_k})$ ;                // representation
        $z_{2k-1} = g(h_{2k-1})$ ;                         // projection
        // second augmentation
        $x_{\tilde{2k}_k} = t'(x_k)$;
        $h_{2k} = f(x_{\tilde{2k}_k})$ ;                    // representation
        $z_{2k} = g(h_{2k})$ ;                             // projection
    end
    foreach $i \in \{1, \dots, 2N\}$ do
        foreach $j \in \{1, \dots, 2N\}$ do
            $s_{i,j} = \frac{z_i^T z_j}{\|z_i\| \|z_j\|}$ ;        // pairwise similarity
        end
    end
    define $l(i,j)$ as: $l(i,j) = -\log \frac{e^{s_{i,j}/\tau}}{\sum_{k=1,k\neq i}^{2N} e^{s_{i,k}/\tau}}$
    $L = \frac{1}{N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)]$;
    update networks $f$ and $g$ to minimize $L$;
end
return encoder network $f(\cdot)$, and discard $g(\cdot)$
```

We follow the original SimCLR[7] setup by defining a multi layer perceptron (MLP) with non-linearities, denoted as $g(\cdot)$, on top of the base CNN, denoted as $f(\cdot)$. Two-layered MLP was defined with ReLU activation in the hidden layer. The hidden dimensions of the MLP were defined as follows 512 input dimensions for the ResNet18, 128 output, following the initial paper implementation (4 times smaller). And for the ResNet50 base 1024 input, 256 output at first. This was changed for the ResNet50 in later experiments to 2048 input and 512 output dimensions, because in the follow-up paper, SimCLRv2[8], the authors mention that wider MLPs boost the performance considerably. The authors also mention that deeper MLPs overfit, so the implementation used only two layers.

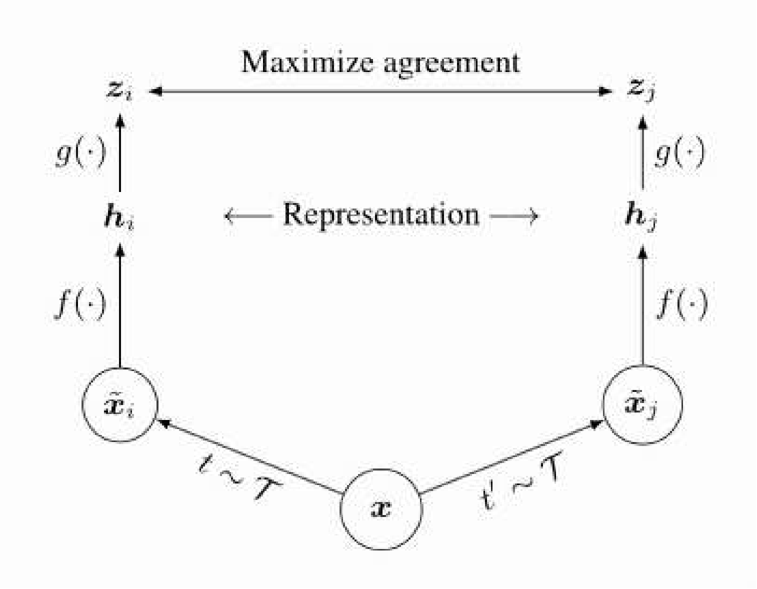


Figure 3.3: SimCLR general setup, Figure credit - Ting Chen et al.[7]

The general setup is visualized in the Figure 3.3.

After the contrastive learning training, shown in algorithm 1, the projection head $g(\cdot)$ is removed, and the base network $f(\cdot)$ is used as a pretrained feature extractor. The representations z that come out of the projection head $g(\cdot)$ have been shown to perform worse[8] than those of the base network $f(\cdot)$ when fine-tuning the network for a new task. This is likely because the representations z are trained to become invariant to many features like the color that can be important for downstream tasks. Thus, $g(\cdot)$ is only needed for the contrastive learning stage.

3.1.3 SimCLR - Training the model

As mentioned before, we want to maximize the similarity between the representations of the two augmented versions of the same image, i.e., z_i and z_j in the Figure shown above 3.3, while minimizing it to all other examples in the batch. SimCLR thereby applies the NT-Xent[8] (Normalized Temperature-scaled Cross Entropy) loss function, a specific implementation of the InfoNCE (Noise-Contrastive Estimation) loss, originally proposed by Gutmann et al. in 2010[26] and later applied to contrastive learning by Aaron van den Oord et al.[44]. In short, the InfoNCE loss compares the similarity of z_i and z_j to the similarity of z_i to any other representation in the batch by performing a softmax over the similarity values. The loss can be formally written as:

$$\ell_{i,j} = \frac{-\log \exp\left(\mathrm{sim}(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}[k \neq i] \exp\left(\mathrm{sim}(z_i, z_k)/\tau\right)} =$$

$$-\mathrm{sim}(z_i, z_j)/\tau + \log\left[\sum_{k=1}^{2N} \mathbf{1}[k \neq i] \exp\left(\mathrm{sim}(z_i, z_k)/\tau\right)\right]$$

The function sim is a similarity metric, and the hyperparameter τ is called temperature, determining how peaked the distribution is. Since the cosine similarity metrics is bounded, the temperature parameter allows us to balance the influence of many dissimilar image patches versus one similar patch. The similarity metric that is used in SimCLR is cosine similarity, as defined below:

$$\mathrm{sim}(z_i, z_j) = \frac{z_i^T \cdot z_j}{||z_i|| \cdot ||z_j||}$$

The maximum cosine similarity possible is 1, while the minimum is -1. In general, the features of two different images converge to a cosine similarity around zero.

ResNet18 and ResNet50 as a base CNNs

The baseline for all experiment results of contrastive learning is, in the beginning, a fully supervised ResNet18, and later a ResNet50 classification of a subset of the HC2022 dataset. The networks are trained from scratch at first on the herbarium subset 200 4.1.3 and carex 4.1.4 datasets, and in later experiments trained on top of the pretrained ImangeNet[18] weights, in order to compare them to their respective SimCLR counterpart architectures (SimCLR experiments were also first ran without the pretrained weights and later with, to document the performance increase). For details on the experiments and the training, please refer to the 4.2.1 Subsection of the results section.

The key characteristic of the ResNet architecture compared to earlier architectures is the introduction of residual blocks. It is stated in[6] "ResNet18 is a 72-layer architecture with 18 deep layers. The architecture of this network aimed at enabling large amounts of convolutional layers to function efficiently. However, the addition of multiple deep layers to a network often results in a degradation of the output. This is known as the problem of vanishing gradient where neural networks, while getting trained through back propagation, rely on the gradient descent, descending the loss function to find the minimizing weights. Due to the presence of multiple layers, the repeated multiplication results in the gradient becoming smaller and smaller thereby "vanishing" leading to a saturation in the network performance or even degrading the performance. The introduction of residual blocks overcomes the problem of vanishing gradient by implementation of skip connections and identity mapping. Identity mapping has no parameters and maps the input to the output, thereby allowing the compression of the network, at first, and then exploring multiple features of the input.”. Residual connections are also a form of regularization (smoothing the error surface).

Visualisation of the ResNet50 architecture, alongside a residual block can be seen in 3.4.

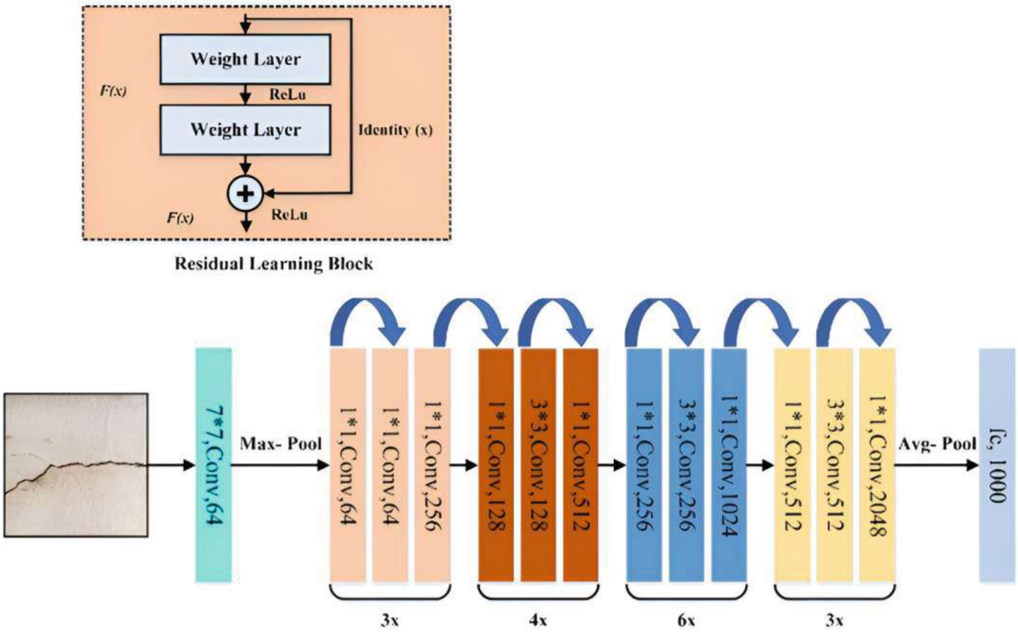


Figure 3.4: ResNet50 Architecture, Figure credit[1]

Logistic regression

After training our model via contrastive learning, it was deployed on a downstream task of plant type classification in order to assess its performance. A common evaluation setup, which also tests the model’s ability to learn generalized representations, involves applying logistic regression[16] to the features. In essence, a single linear layer was trained that maps the representations to class predictions. Since the base network $f(\cdot)$ remains unchanged during the training process, the model’s performance relies on the representations h encapsulating all features necessary for the task. Moreover, overfitting is mitigated, as the model has few trained parameters. Therefore, the anticipation is for the model to perform well even with minimal data.

Logistic regression[16] is a statistical method used for binary classification, predicting the probability of an instance belonging to a particular class. The logistic function, also known as the sigmoid function, transforms a linear combination of input features into a probability score between 0 and 1. The probability $P(y = 1)$ for class 1 is modeled as:

$$P(y=1) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}}$$

Here, b_0 is the bias term, b_i are the coefficients for the input features x_i, and e is the base of the natural logarithm. The log-odds (logit) of the probability is then:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

The model is trained by adjusting the coefficients to maximize the likelihood of the observed class labels in the training data, typically using an optimization algorithm like gradient descent. The decision boundary is determined by the values of the coefficients, separating instances into the two classes based on the calculated probability.

Support Vector Machines

Support Vector Machine (SVM)[15] is a supervised machine learning algorithm that is used in classification and regression problems. SVM tries to find a hyperplane that maximizes the separation between a two-class data set of 2-dimensional space points. The objective is to find a hyperplane that maximizes the separation of the data points to their potential classes in an n-dimensional space. SVM also chooses the hyperplane that separates the data points which has the largest possible margin between the classes. If the data is not linearly separable SVM employs the kernel trick, transforming the data points into a higher dimensional space, depending on the selected kernel function, where a linear separation is possible. The data points with the minimum distance to the hyperplane are called Support Vectors. Only these points have an impact on the final decision boundary.

The computations of data point separation depend on a kernel function. There are different kernel functions: Linear, Polynomial, Gaussian, Radial Basis Function (RBF), and Sigmoid. These functions determine the smoothness and efficiency of class separation. Both RBF and Polynomial Kernels are used in this implementation.

The basic SVM doesn't support multiclass classification. It supports binary classification and separating data points into two classes[15]. The approach used in the thesis is called a "One-to-One" approach. The idea is to map data points to high dimensional space to gain mutual linear separation between every two classes. This "One-to-One" approach breaks down the multiclass problem into multiple binary classification problems. A binary classifier per each pair of classes. In the "One-to-One" approach, the classifier can use $\frac{m(m-1)}{2}$ SVMs.

K-Nearest Neigbours (KNN)

The k-Nearest Neighbors (KNN) algorithm is a supervised learning method used for classification, in the thesis the method works with the image embedding created by the contrastive learning model (512, 2048 dimensional vectors). Euclidean distance is computed between the vectors and used to create a list of nearest neighbors to each vector, after embedding. After that the label or value of a new data point is predicted based on the labels or values of its k closest neighbors in the training set[21].

3.1.4 Segmentation Model for Plants

A YOLOv8[30] segmentation model was trained and utilized to extract only the part of the herbarium image that contains the plant, and another YOLOv8 detection model that detects the rectangular area which contains the plant. Two models are trained in case the model that segments has a lower bounding box detection accuracy than a model that only detects a plant bounding box. In this case to create the dataset that uses the original image background only the detection model is to be used, and for the dataset that replaces the plant background with white colour the segmentation dataset is to be used. This was done in order to reduce the possibility of errors in plant species classification due to the color charts or other visual elements present in the herbarium specimens, other than the plant itself. In the Section 4.1 of the thesis a comparison is shown between a model trained for plant species classification on the specimens that have been segmented from the original image and a model that has been trained on the original images, as well as a model that has been trained on image crops that contain only the plant with its original background. First, a subset of the dataset was selected, and hand annotation was performed on it using the Segment Anything model from META AI[31]. The parts of the herbarium specimen image that contain the pressed and dried plant were selected and labeled to create the dataset. After this a dataset was created and split into proportions of 70% for training, 20% for validation, and the remaining 10% for testing. After that a YOLOv8[30] segmentation model was trained on the dataset. The results of the training, and the subsequent species classification experiments can be found in the final part of the Section 4.1 of the Results.

3.2 Text Recognition

Not all the information is present on all of the labels. The Genus Name and the Species name are present in all of the specimen images in the dataset provided by the NHMW, as described in the datasets Subsection of the results section. Additional data that might be present is the place where the specimen was collected, the name of the person who made the entry, and the date when the herbarium entry was collected. The recognition of the printed and the handwritten label content is evaluated based on state-of-the-art HTR solution, and evaluated on a part of the NHMW herbarium dataset. This is done to show the applicability of text recognition to this type of herbarium specimen analysis. The error rates are calculated based on the recognized text from the plant specimen image labels and the ground truths for the specimens, made available by the NHMW.

3.2.1 Handwritten Text Recognition (HTR)

The emphasis of the HTR is the evaluation of a fully pretrained out-of-the-box approach, Google Vision API, to show the applicability to the herbarium specimen data, more specifically the retrieval of relevant textual information from the specimen images. The solution developed used the Google Vision API. The images were sent to the google vision cloud and the return values were detected text and the regions that hold the text[23], example result can be seen in Figure 3.5.

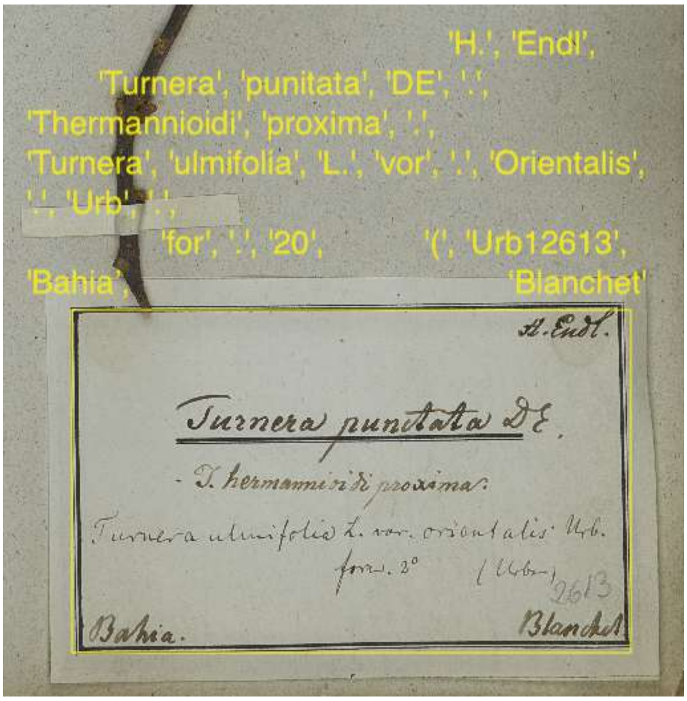

'H.', 'Endl',
'Turnera', 'punitata', 'DE', '.',
'Thermannioidi', 'proxima', '.',
'Turnera', 'ulmifolia', 'L.', 'vor', '.', 'Orientalis',
'.', 'Urb', '.',
'for', '.', '20',
'(', 'Urb12613',
'Bahia',
'Blanchet'

Figure 3.5: Showing an example handwritten label text recognition result from the Goolge Vision Cloud

3.2.2 Optical Character Recognition (OCR)

The Tesseract[35] engine, version 5.3.1, was used for the OCR. The engine supports English and German characters and the out-of-the-box model solution has been used. The used model is the one with LSTM (Long Short Term Memory) layers, for improved word error rate. In order to improve the results of the OCR recognition, and to reduce the noise in the images, the following image preprocessing has been applied: deskewing, grayscaling, noise removal, thresholding, erosion, dilatation and canny filtering. The results were evaluated with and without image preprocessing, the best performing preprocesing methods are selected for the final evaluation that can be found in the results section of the thesis.

CHAPTER 4

Results

The evaluation of the results is split into two different sections. The first section addresses the classification performance of contrastive learning models through linear probing, fine-turning and KNN on the Herbarium 2022[28] dataset. It also assesses the segmentation model for the segmentation of the dried and pressed plant part of the image from the specimens. The species classification performance is also measured on the dataset, created by the aforementioned segmentation model, that only contains the plant parts of the herbarium specimens (evaluating its influence on species classification). The section describes the experiments performed using the ResNet18 and ResNet50 networks, gives the comparison of the fully supervised CNN performance to the contrastive learning models, provides the comparison of the results using different augmentations and training hyperparameters, and finally shows the best performing model classification scores on the family, genus and species level of classification.

The second part addresses the recognition of the printed and the handwritten label content, and shows the performance of OCR and HTR solutions, evaluated on the NHMW herbarium dataset. The error rates are calculated based on the recognized text from the plant specimen image labels and the ground truths for the specimens, made available by the NHMW. All of the datasets used for the experiments are described in the first Subsection of the results.

4.1 Datasets

This section describes all of the datasets used for the experiments of the thesis. There are two main datasets used. The first one consists of 1103 images provided by the NHMW, with their corresponding metadata. The second dataset is the Herbarium Challenge 2022 dataset[28] provided by the New York Botanical Garden. The second dataset was split into two subsets for some of the experiments, both are described in their respective sections.

4.1.1 NHMW dataset

The NHMW has provided an API that holds the ground truth data for some of the specimens in the collection.

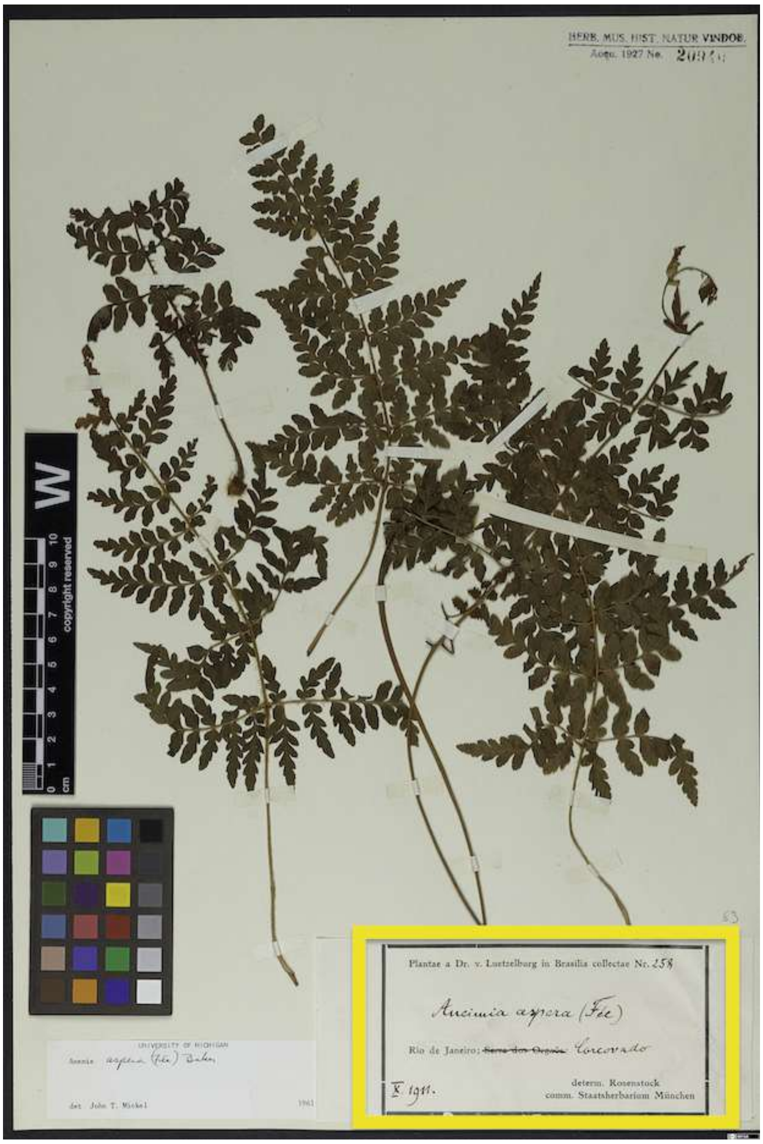
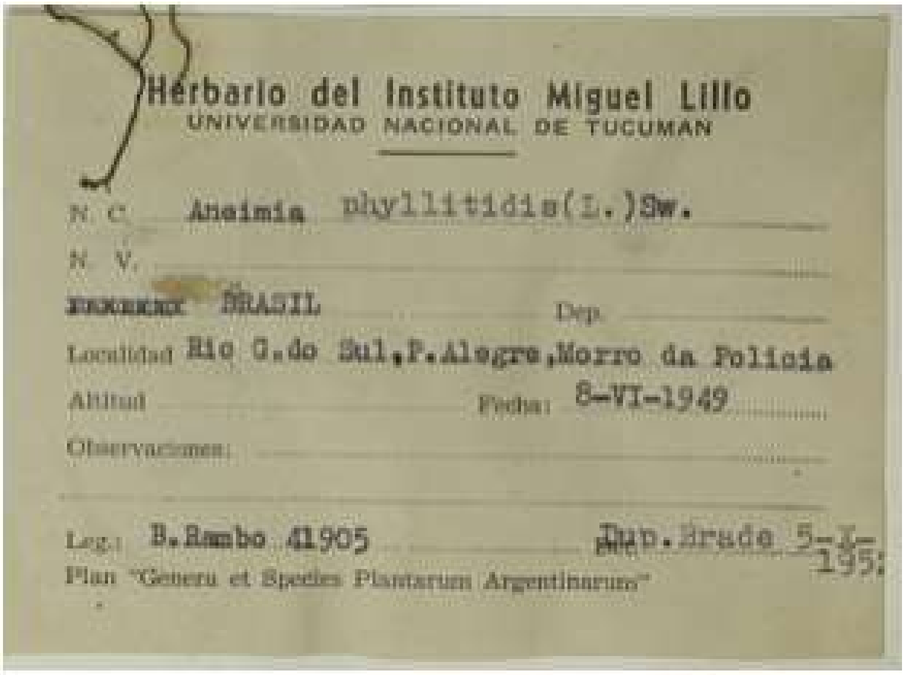

HERB. MUS. HIST. NATUR VINDOB.
Acqu. 1927 No. 2094
copyright reserved
Plantae a Dr. v. Luetzelburg in Brasilia collectae Nr. 258
Anemia aspera (Fée)
Rio de Janeiro: ~~Serra dos Orgãos~~ Corcovado
X 1911.
determ. Rosenstock
comm. Staatsherbarium München

Figure 4.1: Specimen from the NHMW dataset, handwritten label highlighted in yellow

1103 images were provided by the NHMW, with the metadata information for 1078 of them made available, this metadata was retrieved using the API provided by the NHMW. The image resolution is: 10000 x 6660, which makes the file sizes large (10.6GB for 1103 images), this is why we only use this dataset subset, and use other, smaller image size datasets (defined in the next Subsections) for the deep learning experiments. The ground truth and the target class label was also not available for some of the 1103 specimens, making a semi-supervised approach to create a model that does not require target class labels favorable. Creating these ground truths and target class labels also requires large amounts of work and effort from the botanists[5].



Herbario del Instituto Miguel Lillo
UNIVERSIDAD NACIONAL DE TUCUMAN

N. C. Anemia phyllitidis(L.)Sw.
N. V.
[illegible] BRASIL Dep.
Localidad Rio G.do Sul,P.Alegre,Morro da Policia
Altitud Fecha: 8-VI-1949
Observaciones:

Leg.: B.Rambo 41905 Dup.Brade 5-I-195:
Plan "Genera et Species Plantarum Argentinarum"

Figure 4.2: Showing an example of printed information on the label, excerpt from the NHMW dataset

The provided metadata information can be seen in Table 4.1.

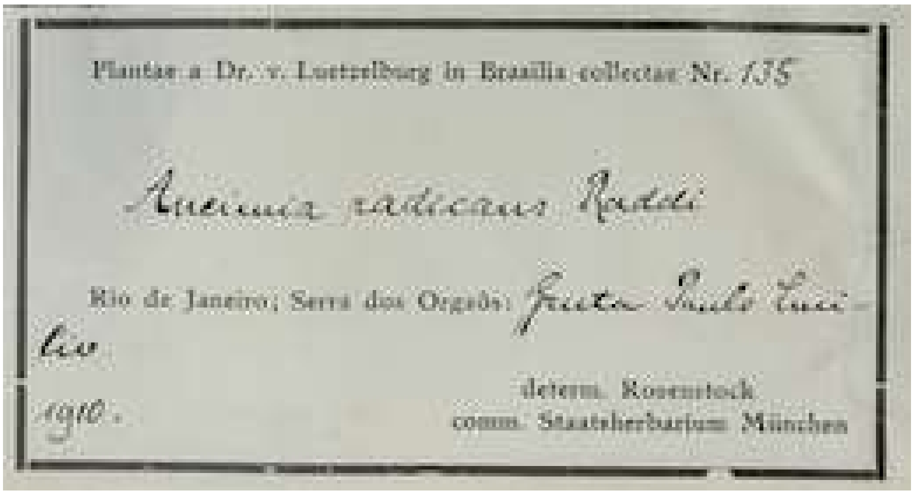
This dataset was used for the evaluation of the OCR and HTR solutions primarily, and for the visualisation application 6 that was developed for the botanical researchers.

4.1.2 Herbarium 2022 challenge dataset[28]

The Herbarium 2022 - FGVC9[28] dataset contains 15,501 species of vascular plants from North America, which constitute more than 90% of the taxa documented in North America, collected across 60 different botanical institutions worldwide. It is a part of a project of the New York Botanical Garden funded by the National Science Foundation to build tools to identify novel plant species around the world. The distribution of class labels across the 1.05M images is long-tailed[28], the distribution of class labels of the training set can be seen in Figure 4.4. There are a minimum of 7 samples per taxa and a maximum of 100 samples per taxa. The maximum number of samples per species in the training set (distribution of quantity shown in Figure 4.4) is 80, others belong to the testing set and the labels of the testing set are not available. Training and test set

Table 4.1: Showing an example specimen metadata from the NHMW dataset, "specificEpithet" holds the plant species name. Family and genus of the plant hold the information relevant to classification as well. The information about the creator and the country of origin of the specimen are also relevant to the botanical researchers. The searchterm field holds the unique name of the file through which the specimen ground truth can be retrieved.

Field	**Value**
searchterm	W19720010868.jp2
title	Anemia phyllitidis (L.) Sw.
description	A PreservedSpecimen of Anemia phyllitidis (L.) Sw. collected by REFLORA provisional entry
creator	REFLORA provisional entry
type	PreservedSpecimen
materialSampleID	https://w.jacq.org/W19720010868
basisOfRecord	PreservedSpecimen
collectionCode	W
catalogNumber	1972-0010868
scientificName	Anemia phyllitidis (L.) Sw.
family	Schizaeaceae
genus	Anemia
specificEpithet	phyllitidis
country	Brazil
countryCode	BRA
recordedBy	REFLORA provisional entry
specimenID	545036
epithet	phyllitidis
collectorTeam	REFLORA provisional entry
OwnerOrganizationAbbrev	W
OwnerLogoURI	http://www.nhm-wien.ac.at/jart/prj3/nhm/resources/images/logo.png
LicenseURI	https://creativecommons.org/licenses/by/4.0/



Plantae a Dr. v. Luetzelburg in Brasilia collectae Nr. 135

Anemia radicans Raddi

Rio de Janeiro; Serra dos Orgaõs: Fruta Paulo Lucio 1910.

determ. Rosenstock
comm. Staatsherbarium München

Figure 4.3: Showing an example of handwritten information on the label, excerpt from the NHMW dataset

ratio is approximately 80%/20% - 839772/210407. The labels of the testing set are not available.

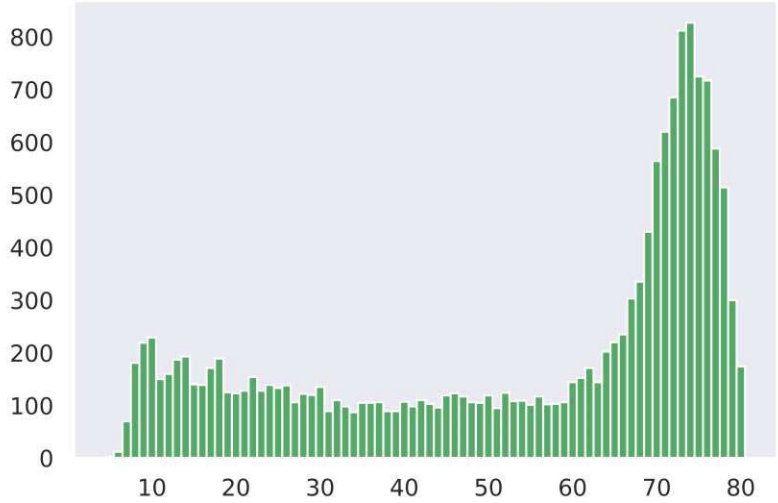


Figure 4.4: Showing the distribution of Herbarium2022 dataset class labels (plant species) by quantity[28], 15501 species in total

In addition to class labels, additional meta-data is available for the family, genus and species. The family and genus level hierarchy has a unique name. There are 272 families and 2564 genera in the Herbarium-2022 dataset. This dataset is skewed at all three levels and has a long-tailed distribution.

The training and test sets contain images of herbarium specimens from 15,501 species of plants. Each image contains exactly one specimen. The text labels on the specimen images have been blurred to remove category information in the image. An example of an image from the dataset can be seen in Figure 4.5.

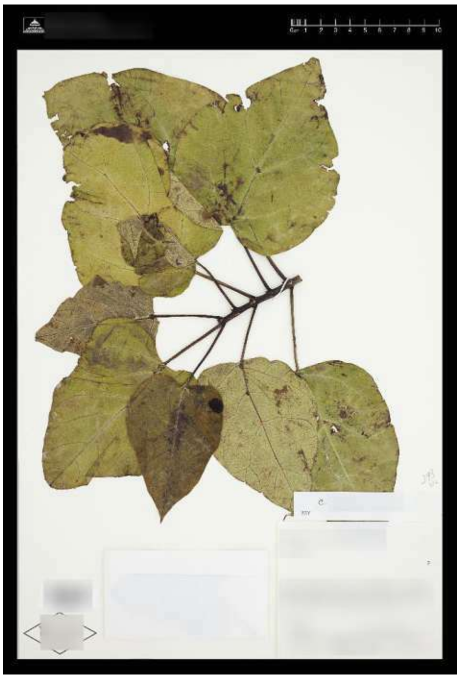


Figure 4.5: Example image from the Herbarium 2022 challenge dataset[28]

The data was approximately split into 80% of the images as the training set and 20% of the images as the testing set. Each category has at least 1 instance in both the training and test datasets. The test set distribution is slightly different from the training set distribution, as mentioned in the documentation, however the labels of the testing set are not available due to the competition guidelines. The training set has a number of examples representing species capped at a maximum of 80.

A hierarchical taxonomic structure of *category_id* representing the plant species is also included in the dataset. The categories in the *training_metadata.json* contain three levels of hierarchical structure, from the highest rank to the lowest rank, namely, family, genus and species. One can think about this as a directed graph, where families are at depth one, genera at depth two, and the species are at depth three. The number of images and classes is summarized as follows:

- Train images: 839772 - Test images : 210407

- Number of classes: 15501 - Image resolution: 1000 x 666

4.1.3 Herbarium 2022 200 biggest categories dataset subset

200 classes of plant species that are the most represented ones within the Herbarium 2022 challenge dataset[28] have been selected as a subset of the full dataset. These 200 classes equate to 15974 images, with their labels. This subset of the dataset was then split: 80% of the images as the training set and 20% of the images as the testing set. All images are from the testing set of the Herbarium 2022 challenge dataset[28]. This dataset was used to test the species classification accuracy of the initial models.

4.1.4 Herbarium 2022 Carex genus dataset subset

The images selected for this dataset were selected, on purpose, to represent the hardest possible classification task for the model: all images are from the same genus, and the goal is to classify them accurately into their correct species. The genus selected from the Herbarium 2022[28] dataset was the one that had the most images (and subsequently different species) associated with it.

The genus selected was Carex, and the number of different species within this genus was 438. The total number of images used in the experiment was: 24743. The images were shuffled at random (same for each experiment).

The dataset split for the Contrastive Learning was: 90% training, 10% validation.

The dataset split used for the logistic regression was: 80% training, 20% testing.

4.1.5 Plant segmentation model dataset

200 images from the Herbarium 2022 dataset were selected, at random, and comprised into a smaller dataset. This dataset is the smallest of the HC 2022 subset datasets, developed to meet the need for manual annotations critical for training a plant segmentation model. Due to the extensive size of other subsets, which precludes feasible hand annotation, this dataset was specifically tailored to support high-quality labeled data for model training through manual segmentation annotations. After that, the areas of these images that contain plants were labeled, by hand, using the Segment Anything model[31] tool, to create annotations for the dataset.

The 200 images were passed through augmentation filters in order to increase the number of samples in the dataset. The filters that were applied: horizontal flip, vertical flip, grayscale (applied to 15% of images), hue (between -15° and +15°), saturation (between -25% and +25%), brightness (between -15% and +15%), exposure (between -10% and + 10%), noise (up to 1.8% of pixels). In total the number of the images in the dataset, including the augmented versions, is 600. The dataset was published and can be accessed at[32].

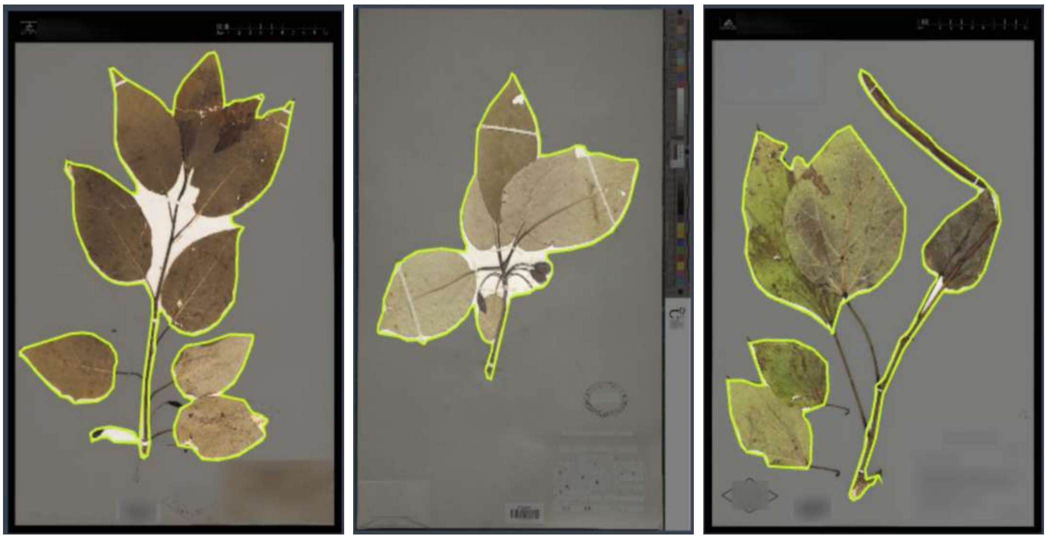


Figure 4.6: Examples from the dataset subset, hand annotated using Segment Anything Model[31]

4.2 Quantitative evaluation of family, genus and species classification

The results of the contrastive learning are the images embedded into a 512, 1024 and 2048 (depending on the CNN model - ResNet18, ResNet50) dimensional vectors that are used to find semantic similarities between the plant species. For the task of species classification, the weights learned in the process of contrastive learning are used to create the embeddings that are used for classification by performing logistic regression over the latent space. The model's evaluation involved comparing species predictions from logistic regression, KNN, and end-to-end fine-tuning of SimCLR model embeddings against true labels to calculate classification accuracy and error rate. This combination of methods provided a comprehensive measure of the embeddings' effectiveness in differentiating species and assessing the model's overall classification performance. The model is evaluated on the Herbarium 2022 dataset[28] and its subsets described in section 4.1.

4.2.1 Fully supervised CNN experiments as baseline

The fully supervised CNN experiments, described in this section, use CNNs, in a fully supervised setting, for the problem of plant species classification. These experiments serve as a baseline to evaluate the effectiveness of the contrastive learning approach. The experiments use datasets described in section 4.1.

This is done because contrastive learning works on top of embeddings created by a convolutional neural network (CNN), as explained in section 3.1, so we can compare the added value of contrastive learning instead of using only a CNN (ResNet18, ResNet50) in a fully supervised setting, trained from scratch in the first experiments and using ImageNet weights in the later experiments.

ResNet18 for the 200 biggest categories dataset

A subset of the dataset was selected, as described in Section 4.1.3, for the experiment, in order to compare the fully supervised classification to the contrastive learning model. These 200 classes equate to 15974 images, all labeled.

The images were resized from the original size (1000x666) into 224x224 pixels (standard input size for the ResNet18 network), and a random grayscale operator was applied with a coefficient of 0.2 (performed on 20% of the training samples,). The application of a random grayscale operator introduces variations in the color space of the images, simulating real-world scenarios where lighting or color conditions may vary. This helps prevent overfitting by encouraging the model to learn more robust and generalized features, rather than relying on specific color cues (camera cues). For further details, refer to "Deep Learning" by Ian Goodfellow et al.[22], specifically the sections discussing data augmentation and robustness enhancement. Batch size was 128. This resizing is a clear disadvantage, since fine-grained details of plants are crucial in species classification. This is a weak point of the comparison and is addressed in later experiments, shown in Subsection 4.2.1.

This subset of the dataset was then split into 80% of the images as the training set and 20% of the images as the testing set. The training of the network was run in 100 and 200 epochs, the network was trained from scratch. The training curve for the accuracy of the model can be seen in Figure 4.7, the curve shows the training and the testing set accuracy, training was performed until convergence in all experiments. All of the fully supervised ResNet experiments use the following hyperparameters: learning rate set to 1e-3, weight decay to 2e-4, batch size to 128, and a 80%-20% training/testing split, following the original SimCLR paper[7]. Results can be seen in the Table 4.2.

Epoch number	Test set accuracy
100 epochs	0.763
200 epochs	0.740

Table 4.2: Inital model accuracy scores on the most represented subset 200 dataset, performance is worse than SimCLR + LogReg

ResNet18 for the Carex genus

An experiment was run on the 438 species of the Carex genus, as described in the Section 4.1.4, using the ResNet18 convolutional neural network in a fully supervised setting. 100 epochs of full supervised training were run. The images were resized from the original size (1000x666) into 224x224 pixels (standard input for the ResNet18 network), and a random grayscale operator was applied with a coefficient of 0.2. Batch size was 128. This

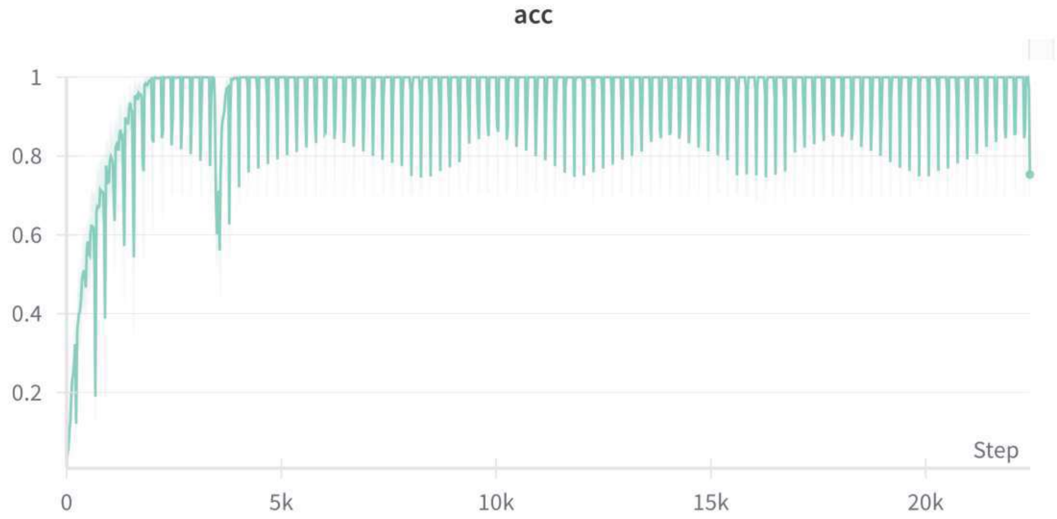


Figure 4.7: ResNet18 learning curve on the 200 categories dataset (drops in the learning curve correspond to the error on the test set, as the accuracy was logged after each loss calculation)

subset of the dataset was then split: 90% of the images as the training set and 10% of the images as the testing set. All of the fully supervised ResNet experiments use the following hyperparameters: learning rate set to 1e-3, weight decay to 2e-4, batch size to 128, and a 80%-20% training/testing split.

The results can be seen in the Table 4.3.

Baseline ResNet50 with ImageNet[18] pretrained weights for the Carex genus

An experiment was run on the 438 species of the Carex genus, using the ResNet50 convolutional neural network in a fully supervised setting with ImageNet pretrained weights as initialization. 200 epochs of full supervised training were run. The images were resized from the original size (1000x666) into 224x224 pixels (standard input for the ResNet18 network), and a random grayscale operator was applied with a coefficient of 0.2. Batch size was 128.

This subset of the dataset was then split: 90% of the images as the training set and 10% of the images as the testing set. All of the fully supervised ResNet experiments use the following hyperparameters: learning rate set to 1e-3, weight decay to 2e-4, batch size to 128, and a 80%-20% training/testing split. The results can be seen in the Table 4.3.

Experiments with random crop selection

The images in the previously described fully supervised CNN experiments were resized into 224x224 pixels, and this resizing is a clear disadvantage, since fine-grained details of plants are crucial in species classification. This is a weak point of the comparison and is addressed in this experiment. The Carex dataset was used for the experiment 4.1.4. All of the parameters for the experiment are the same as in the previous ResNet50 experiment 4.2.1, except that instead of resizing the input images into 224x224 pixels, a random crop of the input image was selected (same as in contrastive learning, but only one crop) of 330x330 pixels, at 20% and the network was trained on these crops. Also 1000 epochs were run, instead of the usual 100 as the network failed to converge in less epochs. The results can be seen in Table 4.3, and the network is 2.3% more accurate than the network in the experiments where the resizing was done.

Experiment	Model	Test accuracy	Train accuracy
Full image input	ResNet18 scratch	0.271	1,0
Full image input	ResNet50 w ImageNet	0.483	1.0
Random crop inputs	ResNet50 w ImageNet	0.506	0.976

Table 4.3: Carex dataset experiment results

4.2.2 Evaluation of SimCLR models using logistic regression

The SimCLR model was trained first on all of the images of the Herbarium 2022 dataset[28] (train + test), using no target class label information, since the model does not require target class labels, all of the available data can be used in a fair setting (for the SimCLR training, not the linear probing that was done after). This model training was then frozen and used to embed the images of a dataset subset, for the subsequent classification (classified using logistic regression over the vector embeddings), to determine the accuracy of the model. Multiple experiments were run to gauge the initial model accuracy, three of which are described in the following Subsection. Each experiment used a different model, trained with different image augmentations as described in Table 4.4. The final model was trained on more than 2 image augmentations per image (for the less represented classes in the dataset).

All of the SimCLR experiments use following hyperparameters, chosen in accordance with the original SimCLR paper[7]: learning rate is 5e-4, temperature set to 0.07, and weight decay is 1e-4, following the initial paper [7].

Linear probing (logistic regression) experiment

In the classification experiments 200 classes of plant species were chosen, as described in Subsection 4.1.3. These 200 classes equate to 15974 images, with their labels. The images were embedded using the contrastive learning model. The embeddings are 512 dimensional vectors, representing their respective image features. A multi class logistic regression model was trained over these embeddings using their target class labels. Adam optimizer was utilized with cross-entropy loss for the training of the logistic regression. All of the logistic regression experiments use the following hyperparameters: learning rate set to 1e-3, weight decay to 1e-3 and the batch size to 64. In the augmentations, first a crop percentage is taken (e.g. 15%), then the crop is resized to desired resolution (e.g. 192x192 pixels).

Experiment results

Model name	Crop resolution	Crop percentage
Model-96-15	96x96 pixels	15% image crop, as in Fig. 4.8
Model-192-35	192x192 pixels	35% image crop, as in Fig. 4.9
Model-192-20	192x192 pixels	20% image crop, as in Fig. 4.10

Table 4.4: SimCLR embedding models augmentation parameters

The resolution in models Model-192-35 and Model-192-20 is the same in order to test whether a higher crop percentage affects the model accuracy positively (20% vs 35%).

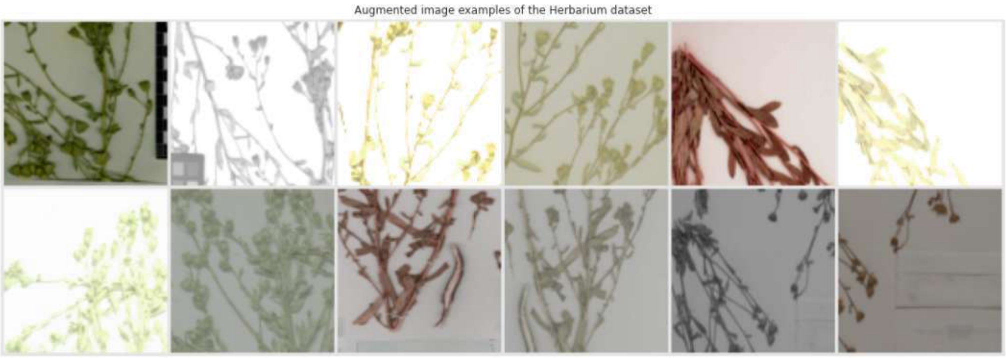


Figure 4.8: Model-96-15: image augmentations at 15% and 96x96 pixels

3000 epochs were run for each logistic regression training (linear probing) and 300 for SimCLR training. The training curve for the accuracy of the logistic regression can be seen in Figure 4.11, the curve shows the training and the testing set accuracy, training was performed until convergence in all experiments. Accuracy scores can be seen in the Table 4.5. The best performing model is the one with the highest resolution and 20% crop percentage (Model-192-20).

4.2.3 Full HC2022 dataset, section 4.1.2, species classification experiment

A linear probing experiment for species classification over the entire dataset was performed. The best performing model (Model-192-20) out of the initial models was used to

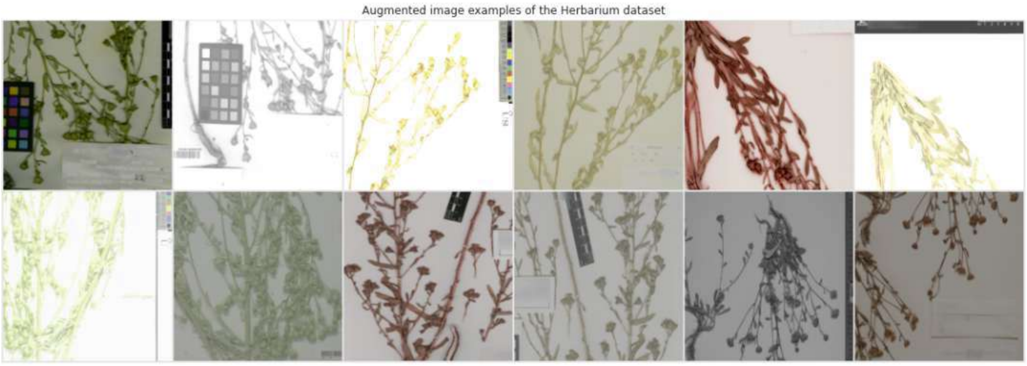


Figure 4.9: Model-192-35 image augmentations at 35% and 192x192 pixels

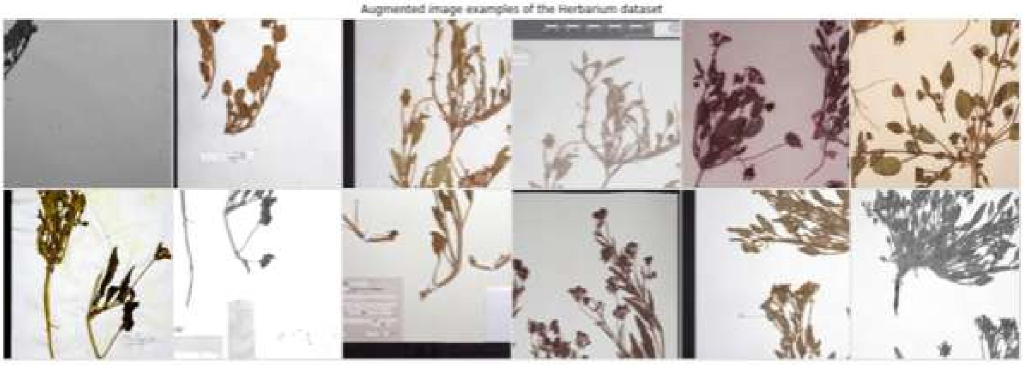


Figure 4.10: Model-192-20 image augmentations at 20% and 192x192 pixels

Model name	Test set accuracy	Train set accuracy
Model-96-15	0.725	0.887
Model-192-35	0.840	0.951
Model-192-20	0.878	1.000

Table 4.5: Initial model accuracy scores on the most represented subset 200 dataset

perform the classification on the full dataset. All of the parameters for the experiment are exactly the same as the Model-192-20 experiment in the previous section, including the augmentations, the logistic regression and the batch sizes. The logistic regression was trained on the 839772 images of the Herbarium 2022 dataset, split 90/10 for training/testing purposes. The model was then used for inference on the 210407 images of the testing set and evaluated on the Kaggle website of the Herbarium 2022 challenge[28].

The accuracy dropped significantly from the initial subset 200 categories dataset experiment accuracy. The species with the most wrong predictions were looked at individually, as shown in Section 4.2.4.

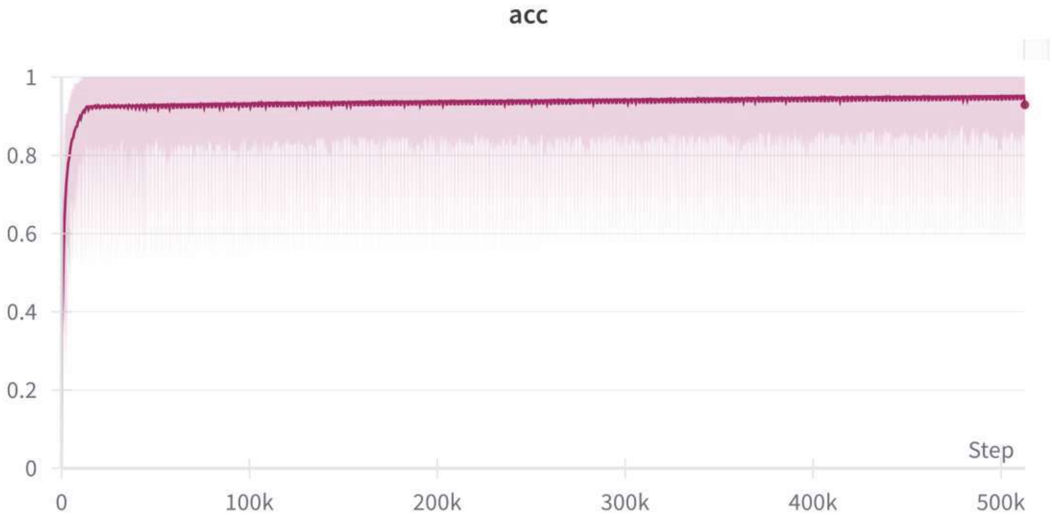


Figure 4.11: ResNet18 learning curve on the 200 categories dataset (drops in the learning curve correspond to the error on the test set)

Model Name	Train Set Accuracy	Test Set Accuracy
Model-192-20	92.27%	52.49%

Table 4.6: Accuracy scores for Model-192-20 on training and test sets at the species level

A full gridsweep of the contrastive learning parameters was done, as described in Section 4.2.6.

4.2.4 Analysis of results based on Hierarchical Plant categorisation

An example of wrong prediction can be seen in Figures 4.12, 4.13, 4.14 and 4.15.

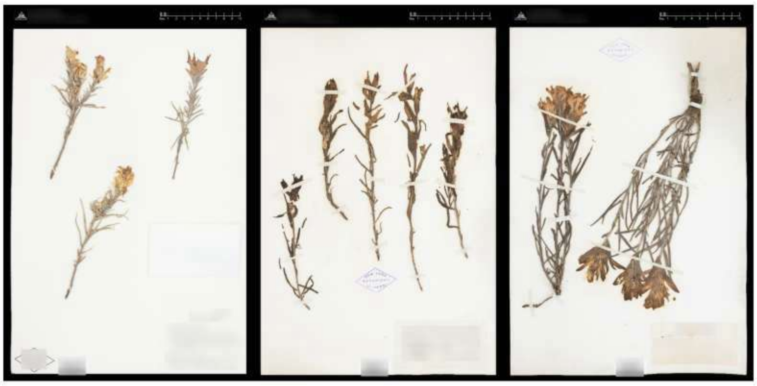


Figure 4.12: Plant for prediction: Orobanchaceae Castilleja integra

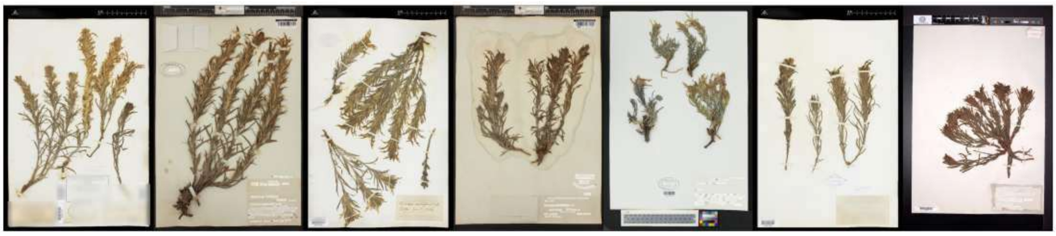


Figure 4.13: Wrong prediction 1: Orobanchaceae Castilleja sessiliflora

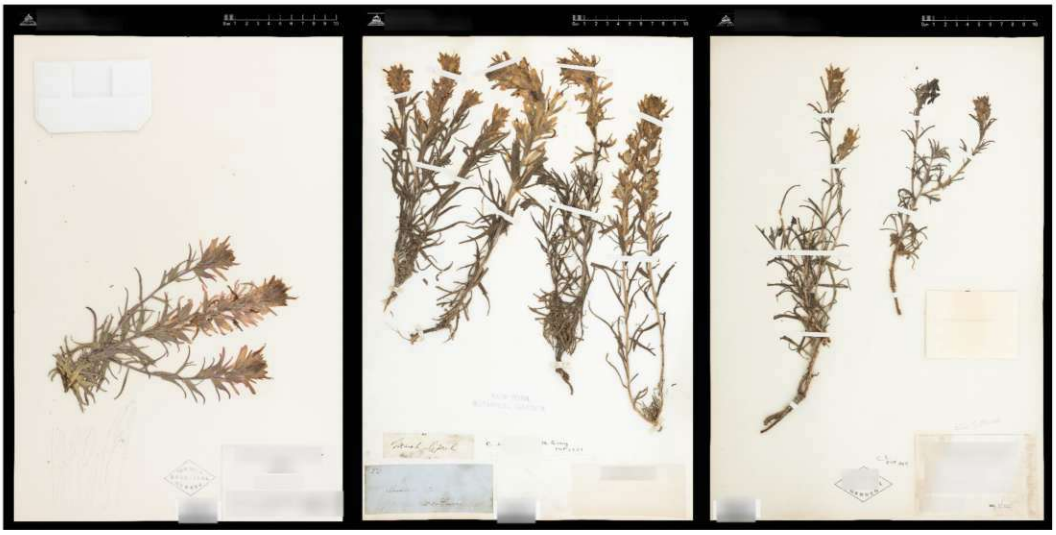


Figure 4.14: Wrong prediction 2: Orobanchaceae Castilleja lindheimeri

To understand the errors, it is essential to first grasp the hierarchical categorization of plants. The classification system is organized as follows:

Plant Family: The highest level of categorization. Plant Genus: A subcategory within the Family. Plant Species: The most specific level within the Genus. This hierarchy is illustrated in Figure 4.16.

Errors in classification can be summarized into three main categories:

- Incorrect Family: This includes errors where both the Genus and Species are also incorrect.
- Correct Family, Incorrect Genus: In this case, the Family is correctly identified, but the Genus and Species are incorrect.
- Correct Family and Genus, Incorrect Species: Here, both the Family and Genus are correctly identified, but the Species is incorrect. Examples of this type of error are shown in Figures 4.12, 4.13, 4.14 and 4.15. Further interpretation of these errors is done in the Section 4.2.5 in the following text.

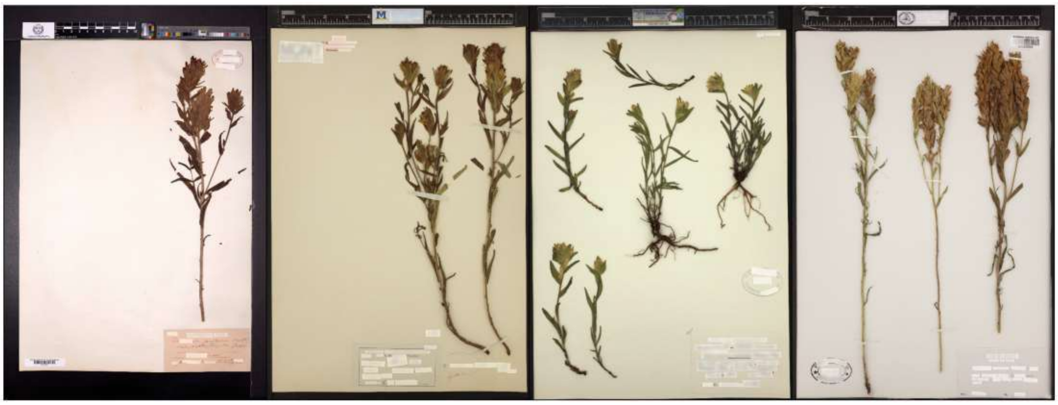


Figure 4.15: Wrong prediction 3: Orobanchaceae Castilleja lutescens

After looking into the results, three different metrics were calculated to gauge the quality of the model. The plants are divided in a hierarchical structure from top to bottom: Family level, Genus level, Species level. The accuracy on each level was calculated and can be seen in Figure 4.16.

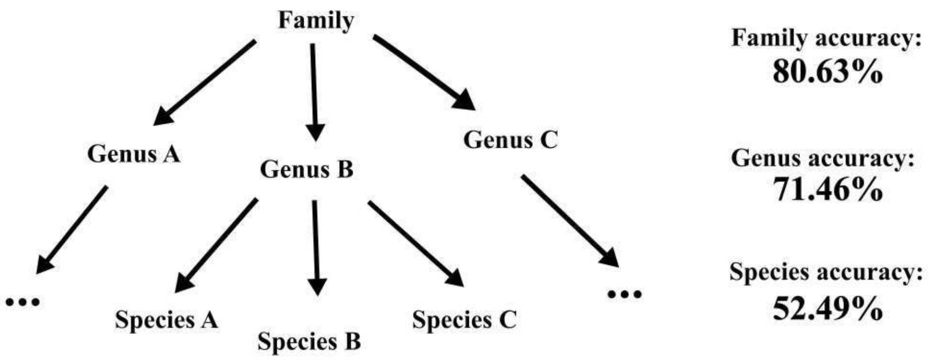


Figure 4.16: Results of Classification on Family, Genus and Species level

4.2.5 Interpretation of the classification results

The mean sample size of species categories in the training set that were correctly classified across all instances was 51.45 images per species. In contrast, the mean sample size of species categories that were misclassified at least once was significantly smaller, averaging 22.29 images per species category. This suggests that more variety and more examples of a species in the training set directly correlates with the model accuracy in species classification. Moreover, the higher accuracy of the Family and Genus level classification, and the lower accuracy of the species level classification highlights the main limitation of the task of plant species classification: the differences between some plants on a species level are extremely detailed and making the species boundary is a difficult problem, even for expert botanists (highlighting the importance of plant species types, described in the introductory section of this thesis)[38], these boundaries are often disagreed upon, and some plants can fall into multiple species categories, in some herbariums, these unresolved or debated taxonomies can result in multiple labels being applied until further research clarifies species boundaries[24]. This also highlights the main advantage of the models that use cropping techniques as they preserve more detailed information compared to using and resizing the entire herbarium specimen image like in the case of fully supervised CNN models that do not use crops.

4.2.6 Gridsweep of image augmentation parameters using SimCLR and Logistic Regression

A gridsweep of the image augmentation parameters used by the contrastive learning model was done, more specifically the resolution and the percentage of the image crops that are considered as a positive pair in the training.

The parameter grid search involved the same two main steps for each parameter pair: the SimCLR training and the logistic regression. The parameters varied during this process included the image crop percentage (starting at 15%) and the image crop resolution (starting at 92x92 pixels).

Due to the computational expense, we used a subset of the dataset for this task, as processing the entire dataset would have been too time-consuming.

The subset of the dataset that used was selected on purpose to represent the hardest possible classification task: all images are from the same Genus, and the goal is to classify them accurately into the correct Species.

The genus selected was *Carex*, as it is the Genus with the greatest number of distinct species, and we are only looking at the species level classification, as discussed in the results interpretation Subsection 4.2.5. The number of Species within this Genus was 438. The total number of images used in the experiment was 24743. The images were shuffled at random (same random seed, set to 42, for each experiment). The dataset is described in section 4.1.4. The dataset split for the Contrastive Learning was 90% training, 10% validation. 100 epochs were run. The batch size for the SimCLR training was 128. The embedding size for the SimCLR output was 512, and the hidden dimension of the projection head was 512 input dimensions and 128 output dimensions, the base network used was ResNet18 with no pretrained weights. The dataset split for the logistic regression was 90% training and 10% testing. 500 epochs were run, the network weights were frozen after the SimCLR training. The batch size for logistic regression was 64. All of the other hyperparameters (learning rate, weight decay and temperature) are the same as in the initial experiments 4.2.2. The crop percentage and resolution range for the experiment can be seen in Figure 4.17.

The results of the gridsweep can be seen in Figures 4.17 and 4.18.

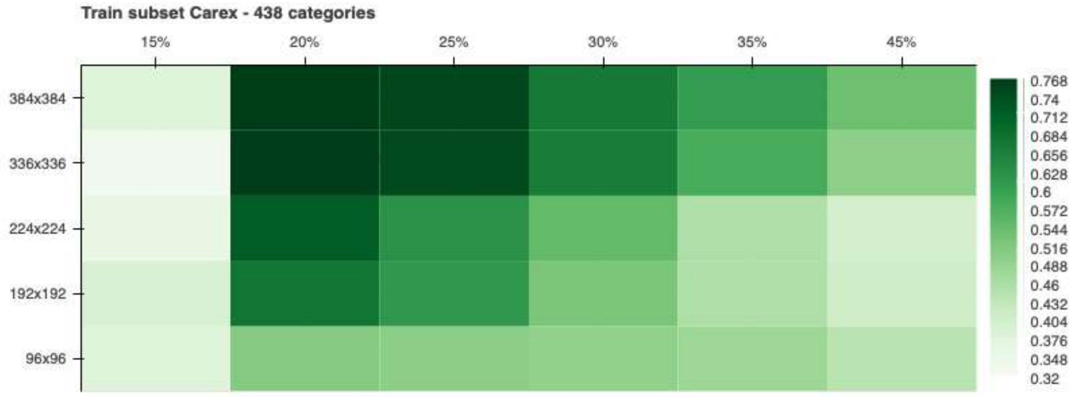


Figure 4.17: Results of Classification Carex, 438 categories, Train set

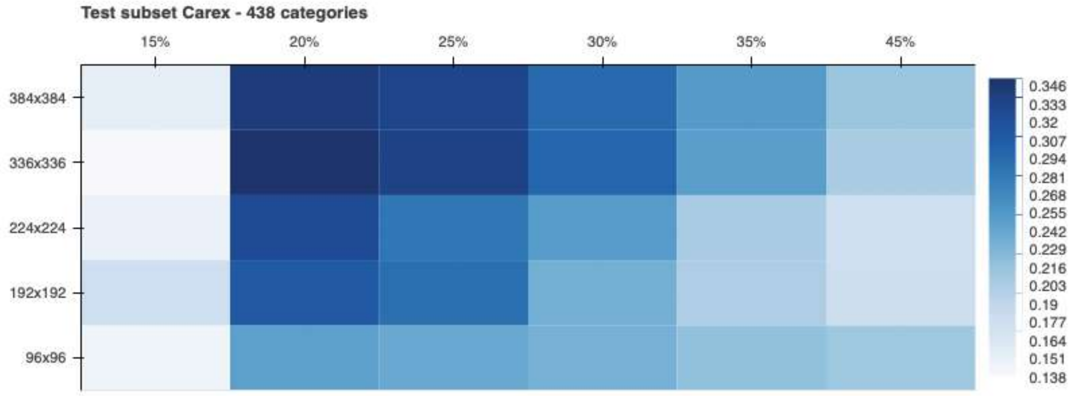


Figure 4.18: Results of Classification Carex, 438 categories, Test set

The best results were given by the 336x336 pixel model at 20% crop.

Accuracy (Species level): 35.01%

4.2.7 ResNet50 as a base for SimCLR training

A larger model was used as the base of the contrastive learning in the experiments. ResNet50 was selected, instead of the ResNet18 used so far and described in Section 4.2.3. The training and the testing of the model was one on the Carex dataset, described in Section 4.1.4. SimCLR training was run and logistic regression on top of it to evaluate the quality of the model. The final layer of the ResNet50 network produces embeddings of 2048 dimensions, instead of the 512 produces by ResNet18. The size of the hidden layer of the projection head (of SimCLR) was set to 512 input and 128 output dimensions at first (same as for ResNet18), however this was changed in further experiments, and is documented in following sections. The dataset split for the Contrastive Learning was 90% training, 10% validation. 100 epochs were run. The batch size for the SimCLR training was 128 (maximum possible due to memory limits of the Vienna Scientific Cluster A100 node). The crop percentage and resolution for the experiment were set to 320x320 pixels at 20% crop, following the gridsweep results described in Section 4.18, and considering the memory limitations of the infrastructure. The dataset split for the logistic regression was 90% training and 10% testing. 500 epochs were run. The batch size for logistic regression was 64. All of the other hyperparameters (learning rate, weight decay and temperature) were set to the same values as in the initial experiments 4.2.2. The experiment results on the Carex dataset can be seen in Table 4.7. An experiment was run, following the same training end evaluation procedure, but using the ImageNet[18] pretrained weights to create the initial embeddings for contrastive learning. The results were superior to the initial model, and can be seen in Table 4.7. Increasing the hidden dimension size, of the projection head to 256 (double than so far), and the first layer of the projection head to 1024 (4*hidden_dimension) improved the results further as seen in Table 4.7.

Base network setup	Test set accuracy	Train set accuracy
ResNet50	0.355	0.863
ResNet50 + ImageNet init	0.474	0.949
ResNet50 + ImageNet init + 2 x hidden_layer	0.545	0.99

Table 4.7: ResNet50 experiment results

4.2.8 SimCLR with ResNet50 with ImageNet pretrained weights base model on the full dataset

A model was trained on the full Herbarium Challenge 2022 dataset, described in Section 4.1.2. It was a SimCLR model, with a ResNet50 model as a base, trained on top of the ImageNet pretrained weights, with the hidden dimension of the projection head set to 256 (1024 input size).

The crop size was 20% at 224x224 pixels (due to performance on such a large dataset, the crop size resolution had to be lowered than in the Carex experiments, to run more epochs on a bigger batch size). The batch size for the experiment was 192 (maximum possible for the hardware used). 100 epochs of SimCLR training were run and 300 epochs of logistic regression on top of the model. The SimCLR training dataset split was 90% training data and 10% validation data. For the logistic regression the split was 90% training data and 10% testing data.

The final test set accuracy on the species level was 0.651, and the results of the species, genus and family level classification accuracy can be seen in Figure 4.19.

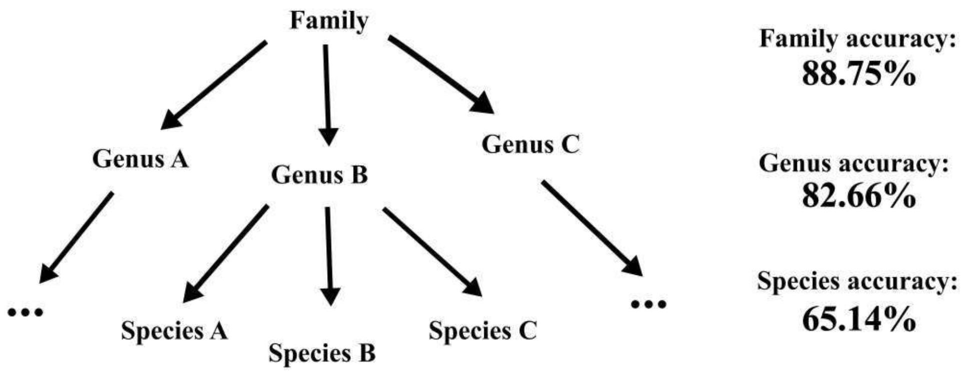


Figure 4.19: Results of Classification on Family, Genus and Species level

4.2.9 Fine-tuning the SimCLR model

Another approach was tested, other than the logistic regression, for the species classification problem, namely end-to-end baseline network fine-tuning. Fine-tuning consists of two parts, first the initial SimCLR training is done in the same self-supervised way as it was done in the experiments described in the previous sections. After that the weights from the base model are copied into another ResNet50 CNN, and a fully connected layer is added on top of them. After copying the weights, the network is trained in a fully supervised end-to-end setting on a dataset, the Carex dataset in this case 4.1.4, as a plant species classifier.

End-to-end fine-tuning experiment setup

The experiments differ in the way that the SimCLR training is performed (the first part of the training), however all of the experiments use exactly the same setup for end-to-end fine-tuning (the second part): the Carex dataset is split into 90% training and 10% testing set for the fine-tuning, 100 epochs of training are run on the baseline CNN (after SimCLR training), the images are all passed without augmentation 1000x666 pixels to the network (same as in every logistic regression experiment) the hyperparameters were the following: batch size 32, learning rate 1e-3, weight decay 2e-4, 100 epochs were run.

Experiment differences

Four experiments were run:

- In the first experiment the Carex genus dataset for SimCLR training was split 90% training and 10% validation. The ResNet50 network loaded with ImageNet pretrained weights was set as the base network. The SimCLR training was done with 20% image crop at 224x224 pixel resolution , the batch size was set to 192 (due to memory limits), and used only the Carex dataset, section 4.1.4. All other parameters were the same as in the initial experiment 4.2.2. The hidden layer of the projection head was set to 256, with the first layer at 4*256 (same as in other ResNet50 experiments). Training was run for 100 epochs. After the training the weights were copied into a new ResNet50 network with a final fully connected layer of 438 output dimensions (number of species in the Carex genus dataset), and 2048 input dimensions (embedding output size of the ResNet50 network). The results can be seen in the Table 4.8 (Experiment 1).

- In the second experiment all of the training parameters and hyperparameters and the architecture setup were the same. The only difference was that the SimCLR network that the weights were copied from was trained on the full Herbarium 2022 dataset 4.1.2, the same as in the full classification experiment 4.2.8. The fine-tuning was still done only on the Carex dataset. The results can be seen in the Table 4.8 (Experiment 2).
- In the third experiment the hidden layer of the projection head of the SimCLR training was increased to 512 dimensions input and 2048 (4*512) dimensions output. The SimCLR training was done on the Carex subset dataset 4.1.4. All other parameters are the same as in the second experiment. The results can be seen in the Table 4.8 (Experiment 3).
- In the fourth experiment the crop size of the contrastive learning model was set to 320x320 at 20% crop - following the best result of the gridsweep experiment 4.2.6, but the number of epochs the model as able to run was 25, significantly less (than the previous model - 100) because of a 128 batch size (maximum due to memory limits). The SimCLR model was trained on the full dataset and the fine-tuning was applied only to the Carex dataset. The model performed worse than the 224x224 one, and the results can be seen in the Table 4.8 (Experiment 4).

Exp. nr.	Crop	SimCLR dataset	hidden layer	Test acc.	Train acc.
1.	224x224	Carex	256	0.727	1.0
2.	224x224	HC2022	256	0.769	0.999
3.	224x224	Carex	512	0.712	1.0
4.	320x320	HC2022	256	0.750	1.0

Table 4.8: Fine-tuning on Carex dataset experiments results

The experimental results indicate that fine-tuning the model outperforms the application of logistic regression as a classifier on its features. When trained exclusively on the Carex dataset, fine-tuning achieves an accuracy of 0.727, compared to 0.545 obtained by logistic regression. Similarly, when trained on the full HC2022 dataset and evaluated solely on the Carex dataset, fine-tuning yields an accuracy of 0.769, whereas logistic regression achieves only 0.554. These findings highlight the effectiveness of fine-tuning in leveraging the model's representational capacity for improved classification performance. The experiments also show that increasing the dimension of the hidden layer of the projection head does not increase the model accuracy.

Full HC2022 dataset end-to-end fine-tuning experiment

In the experiment the HC2022 dataset for SimCLR training was split 90% training and 10% validation. The ResNet50 network loaded with ImageNet pretrained weights was set as the base network. The SimCLR training was done with 20% image crop at 224x224 pixel resolution (following the results from the previous experiments), the batch size was set to 192 (due to memory limits). All other parameters were the same as in the initial experiment 4.2.2. The hidden layer of the projection head was set to 256, with the first layer at 4*256 (same as in other ResNet50 experiments). Training of the SimCLR was run for 100 epochs, maximum allowed running time on the VSC cluster (72 hours). After the training the weights were copied into a new ResNet50 network with a final fully connected layer of 15501 output dimensions (number of species in the HC2022 dataset), and 2048 input dimensions (embedding output size of the ResNet50 network). The results of the classification accuracy on the family, genus and species levels can be seen in the Figure 4.20. The end-to-end fine-tuning was run for 69 epochs, maximum allowed running time on the VSC cluster (72 hours). The training set accuracy was 1.0.

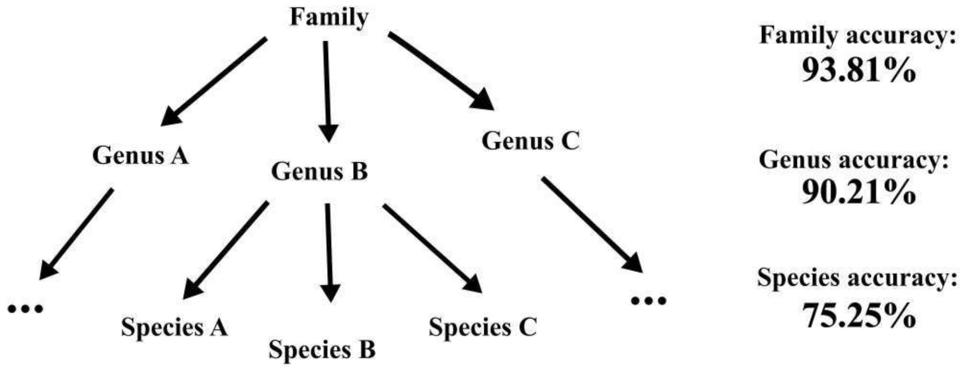


Figure 4.20: End-to-end fine-tuned model classification results on Family, Genus and Species level

4.2.10 Crop Selection based on segmentation

200 images from the Herbarium 2022 dataset were selected, and comprised into a 200 image dataset, as described in the datasets section (segmentation dataset 4.1.5). After that, the areas of these images that contain plants were labeled, manually, using the Segment Anything model[31] tool, to create annotations for the dataset.

The dataset was split into a 70% training set, 20% validation set and a 10% test set. After this, the dataset was used to train a YOLOv8[30] plant detection model and a YOLOv8 plant segmentation model. The results of the training can be seen in Figure 4.21, the test set precision for the segmentation model mask was: 0.871, while the mAP@50 was: 0.895. The mAP@50 of the bounding box finding of the detection model was: 0.995, while the precision was: 1.0. The example of the segmentation inference can be seen in Figure 4.22. A segmentation example can be seen in the Figure 4.22.

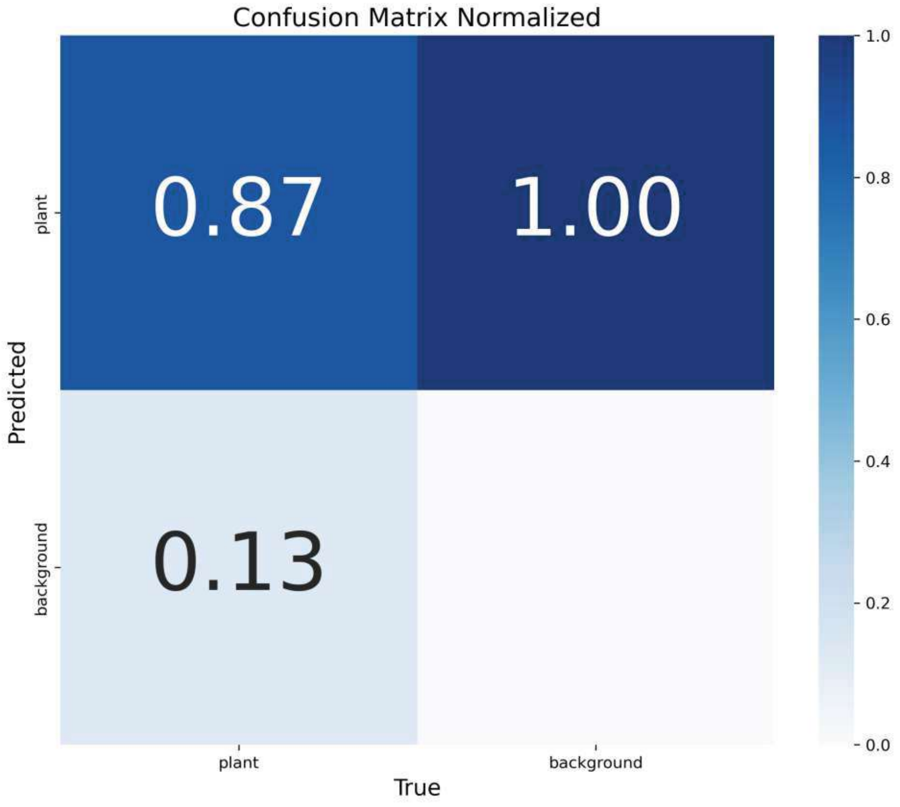


Figure 4.21: YOLOv8 segmentation model confusion matrix

The segmentation model was then used on the 438 species of the Carex genus 4.1.4 to create images that contain only the plant with a black background, and the detection model was used to create a dataset of the same images that contains the original image background but only the area of the plant. Experiments were done using either the entire rectangle area of the image that contains the plant, and another run with just the area of the plant with a black background around it. Thus two separate augmented datasets were created from the "Carex" genus.

The experiments consisted of SimCLR model training and logistic regression classification on top of it, and follow exactly the same hyperparameters and settings as their non-segmented dataset counterparts, explained in Subsections 4.2.7 and 4.2.6. The initial

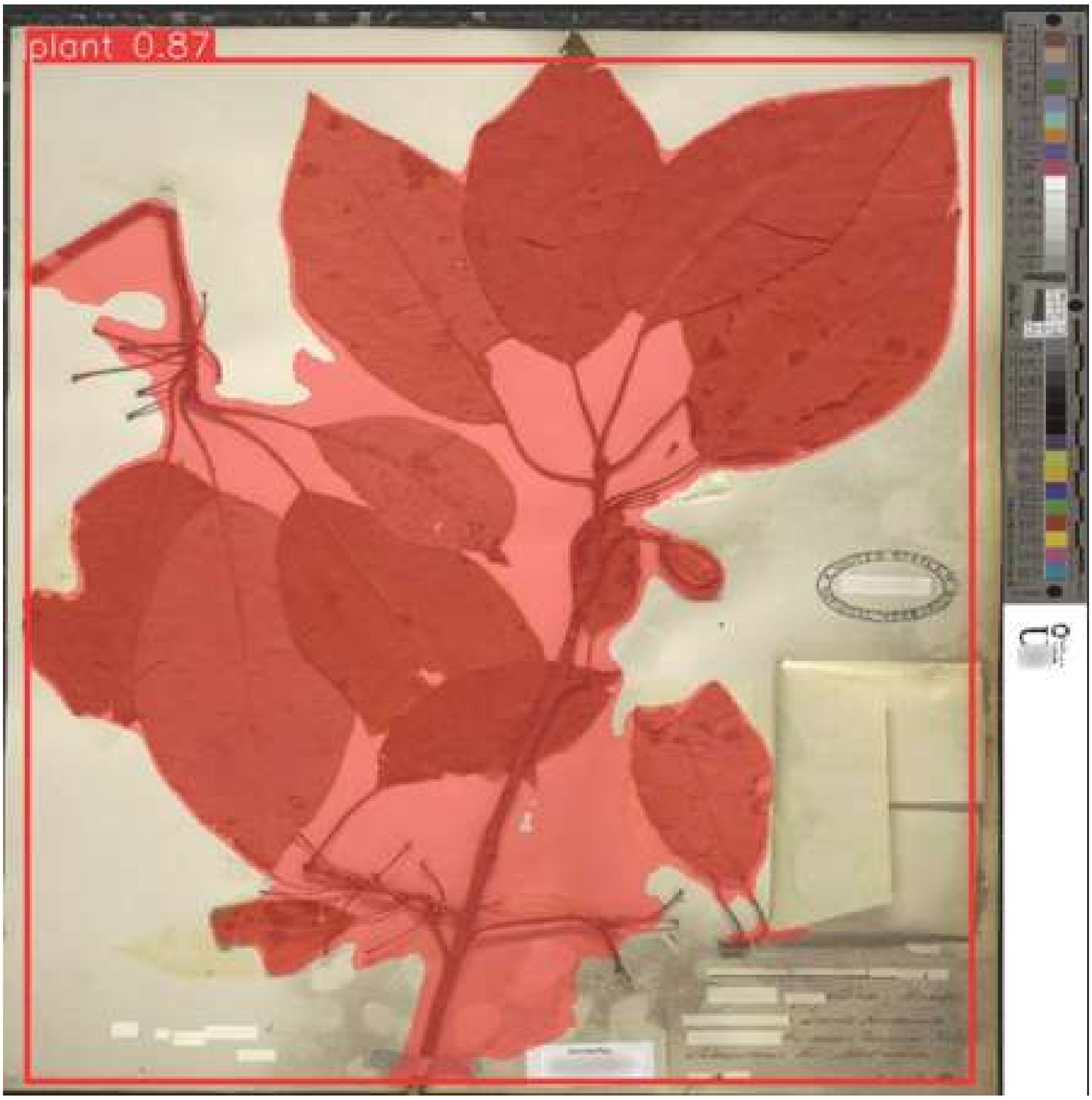


Figure 4.22: YOLOv8 segmentation model result example

results suggest that the method is not outperforming the original method of taking full images, the results can be seen in Table 4.9.

The segmented data showed no increase in the model accuracy. Initial results on the segmented images, and their comparison to the model that uses non-segmented data can be seen in Table 4.9.

Another experiment was conducted using a contrastive learning model that utilized ImageNet[18] pretrained weights as the initial weights, with contrastive learning training subsequently performed on top of these weights, as described in Section 4.2.7, as well as the architecture with increased embedding size (from 512 to 1024). The results 4.10

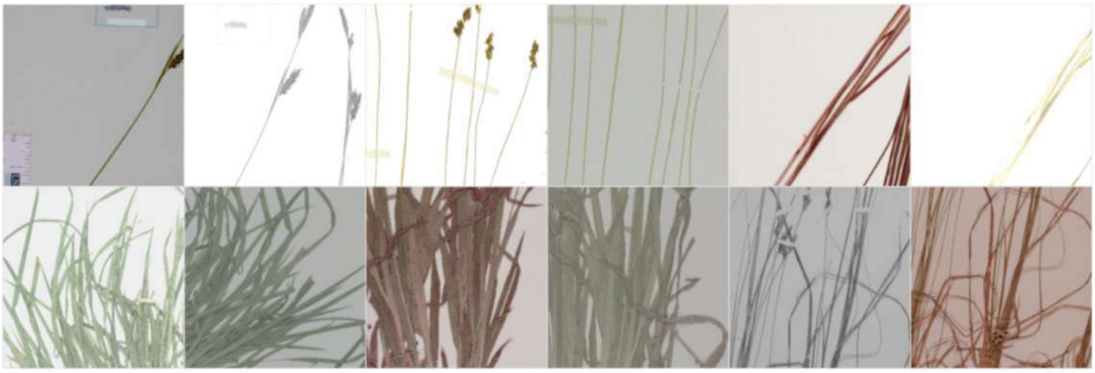


Figure 4.23: Entire rectangle area of the image that contains the plant

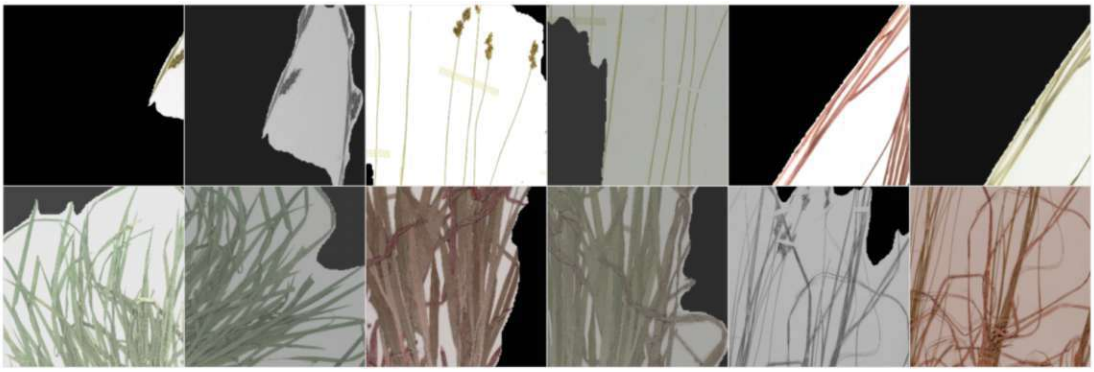


Figure 4.24: Only the area of the plant with a black background around it

Crop Type	Test set accuracy	Train set accuracy
Black background	0.329	0.765
Segmented, original background	0.344	0.703
Original images (not segmented)	0.355	0.863

Table 4.9: Segmented data accuracy scores on the Carex genus dataset, ResNet50 base model

were superior to the initial models, but the segmented data showed not to increase the performance of the model.Increasing the hidden dimension size to 1024 improved the results further as seen in Table 4.11.

The segmented data again showed no increase in the model accuracy (0.544 vs 0.542 for the non-segmented data).

4.2.11 Batch size experiments

In the original SimCLR paper [7] the authors used a batch size of 4096 for their experiments. Due to the memory limitations of the VSC cluster that has been used for the

Segmentation type	Test set accuracy	Train set accuracy
No segmentation	0.474	0.949
Original background	0.444	0.949
Black background	0.425	0.99

Table 4.10: ResNet50,ImageNet weights, 512 hidden dimension size

Segmentation type	Test set accuracy	Train set accuracy
No segmentation	0.544	0.99
Original background	0.542	0.99
Black background	0.484	0.99

Table 4.11: Carex classification accuracy, ResNet50 with ImageNet weights, 128 in 1024 out hidden dimension size.

experiments, the experiments in the thesis have a batch size of 192. In these experiments the batch size was gradually increased to test the effect that the SimCLR training batch size has on the quality of the models. All of the experiments follow exactly the structure of the final ResNet50 experiment (the one using ResNet50 and ImageNet weights as a base) 4.2.7, and use the Carex dataset 4.1.4 in exactly the same split as the gridsweep experiments. The only difference being that we slowly increase batch size until the memory of the computation node allows it. The results can be seen in Table 4.12

Batch size	Crop size	Test set accuracy	Train set accuracy
256	192 x 192 (20%)	0.476	1.0
320	192 x 192 (20%)	0.480	0.998
256	224 x 224 (20%)	0.488	1.0
128	320 x 320 (20%)	0.545	1.0

Table 4.12: Batch size experiment results, Carex dataset

The experiments show that the increase in batch size affects the network accuracy, however it does not affect the accuracy as much as crop size. Bigger crop size yields better performance, and while it shows that a bigger batch size on the 320x320 crops would provide more accuracy, the maximum batch size (128 - due to memory limitations) yields the best model performance.

4.2.12 Other parameter fine-tuning

Removing the color augmentations decreased the performance as well. Changing grayscale, hue and saturation augmentations from 0.5 decreased the performance as well. Changing the brightness of the augmentations decreases the performance. Varying the augmentation crop size in the experiments decreases the performance as well.

4.2.13 Additional experiments

K-Nearest Neighbours

The best trained SimCLR model was loaded, as explained in section 4.2.8. The carex dataset 4.1.4 (all specimens from the same genus) images were passed through the model to create 2048 dimensional embeddings. KNN was run to find the 7 nearest neighbours to each image. The results of the top 1 same species accuracy and the top 6 species accuracy can be seen in Table 4.13.

Dataset	Top 1 accuracy	Top 6 accuracy
Carex 4.1.4	0.118	0.283

Table 4.13: KNN on top of SimCLR accuracy scores

Multiclass Support Vector Machine

In the experiment, 200 classes of plant types that are the most represented ones within the Herbarium 2022 dataset have been selected, as shown in Section 4.1.3. The images were embedded using the contrastive learning model. The embeddings are 512 dimensional vectors, representing their respective image features. A multi-class SVM classifier was run over these image embeddings, using two different kernels: a polynomial kernel, and an RBF kernel.

The experiment results can be seen in the Table 4.14.

Kernel type	Accuracy	F1 score
Polynomial Kernel	0.639	0.663
RBF Kernel	0.181	0.163

Table 4.14: SVM experiment results on the biggest 200 categories dataset (see Section 4.1.3)

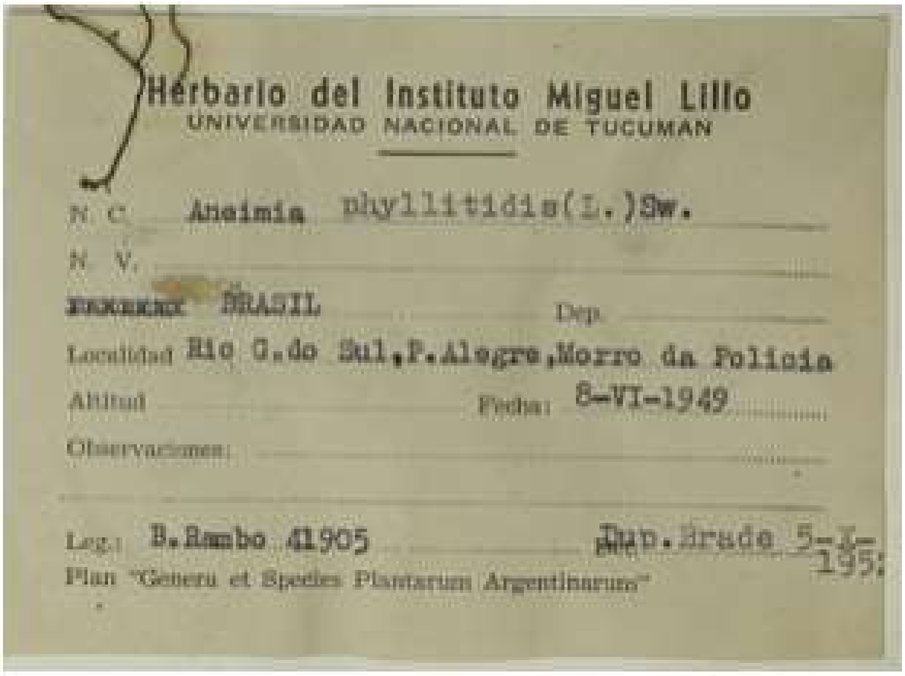
The RBF kernel overfit, and the Polynomial kernel has given a better result, however the results show that the SVM produces a lower score than the logistic regression, so it was decided that the further experiments focus on the logistic regression method.

4.3 Character Recognition Evaluation

The outputs of our text recognition solution (see Methodology) have been evaluated on the dataset of herbarium specimen made available by the NHMW, shown in Subsection 4.1.1. All of the ground truth labels contain the plant genus, type and species, so only that information is compared, as it is present in all of the specimen images. The results of the HTR solution are compared with the OCR Tesseract engine for the printed labels, to see if an HTR model is necessary as a better solution.

4.3.1 OCR Results

OCR was done using the Tesseract engine. Some images have printed information, some have handwritten information. An example of the printed information on the label can be seen in the Figure 4.11. The full label of the plant described by the label in Figure 4.11 is "Aneimia phylitidis (L.) Sw.".



Herbario del Instituto Miguel Lillo
UNIVERSIDAD NACIONAL DE TUCUMAN

N. C. Aneimia phyllitidis(L.)Sw.
N. V.
BRASIL Dep.
Localidad Rio G.do Sul,P.Alegre,Morro da Policia
Altitud Fecha: 8-VI-1949
Observaciones:

Leg.: B.Rambo 41905 Dup.Brade 5-X-195
Plan "Genera et Species Plantarum Argentinarum"

Figure 4.25: Example of printed information on the label

Ten images with printed information were selected by hand for the initial evaluation. The ground truth information of the plant genus and species was searched for in the recognized text. The full plant name was searched for first in the recognized text (full term error rate).

Experiment description	Accuracy
All words error rate	0.050
Full plant name accuracy	0.700

Table 4.15: OCR 10 selected images with printed information

1078 images were selected at random from the NHMW dataset for the secondary evaluation, as detailed in Section 4.1.1.

Looking at them by hand, 344 (31.91%) of the images in this set have printed information on them, the others have handwritten labels.

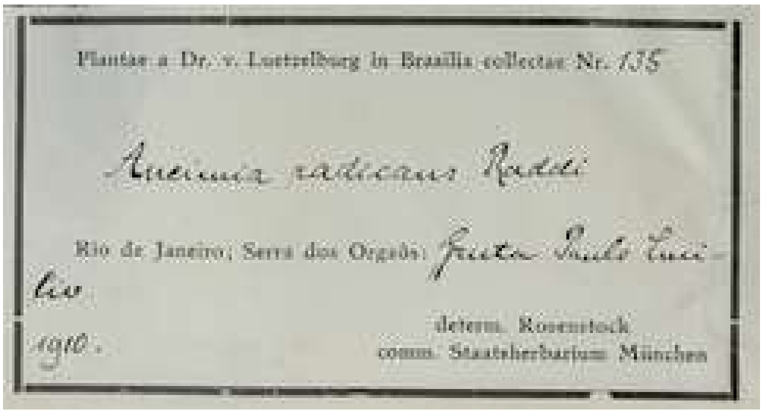
Experiment description	Accuracy
Shortened plant name accuracy (Genus + Species)	0.213
Plant Genus accuracy	0.286
Full plant name accuracy (with end abbreviations)	0.089

Table 4.16: OCR results on all 1078 images

Out of all the words that include: Plant Family, Plant Genus or Plant Species within these 1078 images OCR has correctly detected: 20.54% of them. Making the word error rate: 79.36% (on all images of the NHWM dataset, handwritten labels included).

4.3.2 HTR Results

HTR evaluation was done using the Google Cloud Vision HTR engine. Some images have printed information, some have handwritten information. An example of the printed information on the label can be seen in the Figure 4.12. The full label of the plant described by the label in Figure 4.12 is "Aneimia radiccans Raddi".



Plantae a Dr. v. Luetzelburg in Brasilia collectae Nr. 135

Aneimia radicans Raddi

Rio de Janeiro; Serra dos Orgaõs: Fructa Paulo Luci-
lio

determ. Rosenstock
comm. Staatsherbarium München

1910.

Figure 4.26: Example of handwritten information on the label

The same 1078 images selected from the NHMW Herbarium used in the OCR evaluation were used in the HTR evaluation as well.

Experiment description	Accuracy
Shortened plant name accuracy (Genus + Species):	0.391
Plant Genus accuracy:	0.600
Full plant name accuracy (with end abbreviations):	0.039

Table 4.17: Inital model accuracy scores on the most represented subset 200 dataset

Total word error rate was: 65.81%. 34.19% words were detected in total.

CHAPTER 5

Conclusion

Throughout this thesis, comprehensive analyses and experiments were conducted to investigate the following research questions:

- RQ1: What is a suitable deep learning architecture for the analysis of the herbarium plant specimens?
- RQ2: How can state-of-the-art text recognition (handwritten and printed) methods be used for herbarium data analysis?
- RQ3: How well do self-supervised contrastive learning methods perform on the task of plant species classification?

The research questions can be answered as follows:

What is a suitable deep learning architecture for the analysis of the herbarium plant specimens?

The thesis demonstrates that a combination of deep convolutional neural networks (CNNs) with a self-supervised learning approach like SimCLR provides a suitable architecture for analyzing herbarium specimens. Models such as ResNet18 and ResNet50 were used as backbone networks for SimCLR training, and feature extraction. The experiments showed that embedding representations derived through SimCLR, followed by end-to-end fine-tuning and linear probing (e.g., logistic regression), achieved higher classification accuracy compared to fully supervised CNNs, and also proved to be suitable in scenarios with limited labeled data by employing KNN. The results suggest that SimCLR-based architectures, augmented by proper image cropping and segmentation techniques, are effective for extracting semantically rich features from herbarium specimen images. Additionally, segmentation models like YOLOv8 were used to focus on plant features by isolating plant regions from background elements, further improving classification performance.

This thesis demonstrates that deep learning can significantly alleviate the labor-intensive task of manual plant classification by botanists, particularly at the species level, where differences are often extremely detailed. Establishing species boundaries is a challenging problem, even for experts, underscoring the critical role of "plant species types" as reference points for accurate classification [38]. Moreover, these boundaries are frequently debated, and some plants fall into multiple species categories due to unresolved or evolving taxonomies, often resulting in provisional labels being applied until further research provides clarity [24]. By leveraging deep learning techniques, such as self-supervised contrastive learning and fine-tuning, this thesis offers automated solutions that reduce the manual workload while enabling consistent, scalable, and robust taxonomic analysis, even in the face of such complexities.

Beyond image-based classification, the thesis highlights the integration of Handwritten Text Recognition (HTR) and Optical Character Recognition (OCR) systems to analyze textual data found on specimen labels. These labels contain vital botanical information, including species, genus, family, collection location, collector names, and collection dates. By extracting this information, HTR and OCR systems provide crucial metadata that complements image analysis.

For example:

- Species Identification: Transcribed genus and species names from labels can be cross-referenced with classification models to validate or enhance predictions.
- Historical and Geographical Studies: Collection dates and locations retrieved from labels enable researchers to analyze phenological changes over time and assess geographical distribution patterns.
- Data Digitization: Automating the transcription of handwritten and printed text reduces manual effort and accelerates the digitization of large herbarium collections.
- Collaborative Research: Extracted metadata facilitates integration with external botanical databases, fostering collaborative research and enabling global studies.

This dual approach of combining visual and textual data analysis enhances the accuracy, scope, and utility of herbarium data processing, creating a robust framework for botanical research. The integration of these systems into a single analysis pipeline makes the architecture not only suitable but transformative for herbarium specimen analysis.

Finally, the thesis describes the development of a visualization application (detailed in the appendix) to support botanical researchers. This application leverages the embeddings generated by the SimCLR framework to create an interactive platform for exploring clusters of similar plant specimens. Key features of the application include:

- Cluster Exploration: Researchers can navigate groups of similar specimens based on visual and textual features, providing insights into species relationships and patterns.
- Detailed Inspection: Individual specimens can be examined in greater detail, with access to both image data and associated metadata.
- Enhanced Search Tools: The application integrates search functionalities to quickly locate specimens within large datasets, facilitating efficient retrieval of relevant information.

By integrating visual and textual data analysis within a user-friendly interface, this application enhances the utility of the developed deep learning architecture and demonstrates its practical applicability in real-world botanical research.

How can state-of-the-art text recognition (handwritten and printed) methods be used for herbarium data analysis?

State-of-the-art Handwritten Text Recognition (HTR) and Optical Character Recognition (OCR) systems, including the Google Vision API and Tesseract OCR, were evaluated for their ability to extract text from herbarium labels. These systems demonstrated strong performance in recognizing genus, species, collection locations, and collector names, especially when coupled with preprocessing techniques like greyscaling and noise reduction. The recognition of both printed and handwritten text was feasible, providing an effective means to digitize herbarium metadata. These methods can be integrated with deep learning-based specimen analysis tools, allowing for a more automated and scalable approach to metadata transcription and enhancing accessibility to digitized herbarium collections. The experiments of the text recognition have shown that the HTR method performed better than the OCR method, with an accuracy of the shortened plant name recognition of: 0.391. This was done with an out-of-the-box solution with no additional training on the herbarium data.

How well do self-supervised contrastive learning methods perform on the task of plant species classification?

The supervised deep learning (contrastive learning) experiments that employ end-to-end fine-tuning on top of contrastive learning achieved the following accuracy scores on the HC2022 dataset 4.1.2: species accuracy: 0.752, genus accuracy: 0.902 , family accuracy: 0.938. The supervised deep learning (contrastive learning) experiments with linear probing have yielded the following accuracy scores on the HC2022 dataset 4.1.2: species accuracy: 0.651, genus accuracy: 0.826 , family accuracy: 0.888. And the KNN on top of the self-supervised learning has achieved an accuracy of 0.288 compared to 0.551 of contrastive learning with linear probing and 0.471 of supervised CNN deep learning, and 0.726 of contrastive learning with fine-tuning. This shows that unsupervised techniques are valuable to those datasets that do not have labels. SimCLR has also outperformed a fully supervised CNN on the Carex dataset 4.1.4, with a score of:0.769 (end-to-end fine-tuning) , and a score of: 0.545 (logistic regression) compared to the CNN score of: 0.483. The higher accuracy of the Family and Genus level classification, and the lower accuracy of the species level classification highlights the main limitation of the task of plant species classification: the differences between some plants on a species level are extremely detailed and making the species boundary is a difficult problem, even for expert botanists (highlighting the importance of plant species types, described in the introductory section of this thesis)[38]. This also highlights the main advantage of the models that use cropping techniques as they preserve more detailed information compared to using and resizing the entire herbarium specimen image like in the case of fully supervised CNN models that do not use crops. The findings confirm that self-supervised learning combined with fine-tuning offers a powerful solution for plant species classification, outperforming traditional supervised CNN approaches and linear probing. This methodology provides an accurate and efficient framework for herbarium specimen analysis, with significant implications for automating botanical research and enhancing access to digitized collections.

CHAPTER 6

Future Work

Deep learning of plant phenology

Building on the findings of this thesis, future research on the deep learning of plant phenology should explore advanced self-supervised learning architectures that have surpassed SimCLR in performance. Since the release of SimCLR v2 [8], more recent contrastive learning methods, such as MoCoV3 [27], have demonstrated superior results. Additionally, the emergence of self-supervised vision transformer architectures [43] has pushed the boundaries of representation learning further, achieving remarkable advancements in downstream tasks.

Despite the availability of newer methods, SimCLR was chosen for this thesis due to its well-documented nature, comprehensive resources, and ease of implementation, which provided a practical foundation for experimentation. However, to advance the application of self-supervised learning in plant phenology analysis, the performance of MoCoV3 and self-supervised vision transformers should be evaluated on similar tasks, such as clustering and classifying phenological traits across plant species.

Adopting these methods may require addressing technical challenges, particularly the computational demands inherent in contrastive learning. For instance, batch size plays a critical role in determining the quality of learned representations, as contrastive learning depends on comparing each augmented image pair against all other pairs within a batch. This scaling challenge highlights the necessity of powerful computational infrastructure to fully utilize the potential of advanced architectures like MoCoV3 and vision transformers.

Furthermore, expanding the scope of phenology analysis to include temporal and environmental data could provide a richer context for understanding plant lifecycle events such as flowering and seed dispersal. Integrating such multimodal data with more advanced self-supervised methods would allow for more nuanced insights and further enhance the applicability of deep learning in phenological studies.

Visual analysis application for the NHM

Adding more models to choose from is another suggestion for the future of the application, training and adding models like the vision transformer model[43], or the MoCoV3 model[27] should provide diversity of choice for the researchers, and lead to new knowledge gained based on the selected model clustering. Creating pan and zoom options, for the visualisation tool, to browse through the latent space by reducing its dimensions (through PCA or UMAP) is also a feature that could benefit the botanical researchers. The application should allow the botanical researchers to look through more similar specimen than the current ten, that the application is limited to. A qualitative analysis of the application can be conducted to evaluate what information the botanical researchers need (and find most useful) from the deep learning models.

Text Recognition of the herbarium specimen

A state-of-the-art solution in HTR, such as DAN[11], should be applied and trained successfully to the NHMW herbarium dataset in order to increase the accuracy and help the information retrieval process, as the annotations made by botanists hold important information. The scope of this thesis has limited the research in this segment of herbarium analysis. It should be noted that the plant classification is an instance segmentation task in itself for the researchers, and it can happen that the boundaries between the species are not clear, as one species slowly transitions into another species[42], thus the handwritten information retrieval should definitely not be used as the only means of herbarium digitalisation.

List of Figures

List of Tables

Bibliography

[1] L. Ali, F. Alnajjar, H. Jassmi, M. Gochoo, W. Khan, and M. Serhani. Performance evaluation of deep cnn-based crack detection and localization techniques for concrete structures. *Sensors*, 21:1688, 03 2021.

[2] Anno-AI. Evaluating offline handwritten text recognition: Which machine learning model is the winner?, 2020. Available at: https://anno-ai.medium.com/part-2-evaluating-offline-handwritten-text-recognition/-which-machine-learning-model-is-the-e8c392b76328.

[3] D. P. Bebber, M. A. Carine, J. R. I. Wood, A. H. Wortley, D. J. Harris, G. T. Prance, G. Davidse, J. Paige, T. D. Pennington, N. K. B. Robson, and R. W. Scotland. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences*, 107(51):22169–22171, 2010.

[4] G. Besnard, M. Gaudeul, S. Lavergne, S. Muller, G. Rouhan, A. Sukhorukov, A. Vanderpoorten, and F. Jabbour. Herbarium-based science in the twenty-first century. *Botany Letters*, 165:323–327, 10 2018.

[5] J. Carranza-Rojas, H. Goëau, P. Bonnet, E. Mata-Montero, and A. Joly. Going deeper in the automated identification of herbarium specimens. *BMC Evolutionary Biology*, 17, 12 2017.

[6] Y. Chandola, J. Virmani, H. Bhadauria, and P. Kumar. Chapter 4 - end-to-end pre-trained cnn-based computer-aided classification system design for chest radiographs. In Y. Chandola, J. Virmani, H. Bhadauria, and P. Kumar, editors, *Deep Learning for Chest Radiographs*, Primers in Biomedical Imaging Devices and Systems, pages 117–140. Academic Press, 2021.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[8] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[9] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, October 2021.

[10] S. Chulif and Y. L. Chang. Herbarium-field triplet network for cross-domain plant identification. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*, page 173–188, Berlin, Heidelberg, 2021. Springer-Verlag.

[11] D. Coquenet, C. Chatelain, and T. Paquet. Dan: A segmentation-free document attention network for handwritten document recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8227–8243, July 2023.

[12] D. Coquenet, C. Chatelain, and T. Paquet. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):508–524, jan 2023.

[13] D. Corney, J. Clark, H. Tang, and P. Wilkin. Automatic extraction of leaf characters from herbarium specimens. *Taxon*, 61:231–244, 02 2012.

[14] D. P. A. Corney, H. L. Tang, J. Y. Clark, Y. Hu, and J. Jin. Automating digital leaf measurement: The tooth, the whole tooth, and nothing but the tooth. *PLOS ONE*, 7(8):1–10, 08 2012.

[15] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[16] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[17] R. de Lutio, J. Y. Park, K. A. Watson, S. D'Aronco, J. D. Wegner, J. J. Wieringa, M. Tulig, R. L. Pyle, T. J. Gallaher, G. Brown, G. Guymer, A. Franks, D. Ranatunga, Y. Baba, S. J. Belongie, F. A. Michelangeli, B. A. Ambrose, and D. P. Little. The herbarium 2021 half–earth challenge dataset and machine learning competition. *Frontiers in Plant Science*, 12, 2022.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[20] E. Farnsworth, M. Chu, J. Kress, A. Neill, J. Best, J. Pickering, R. Stevenson, G. Courtney, J. Dyk, and A. Ellison. Next-generation field guides. *BioScience*, 63:891–899, 11 2013.

[21] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989.

[22] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[23] Google LLC. Google cloud vision, 2024. Available at: https://cloud.google.com/vision.

[24] V. Grant. *Plant Speciation*. Columbia University Press, 1981.

[25] P. J. Grother. Nist special database 19. nist handprinted forms and characters database, 2008-10-16 14:10:54 2008. Available at: `https://www.nist.gov/srd/nist-special-database-19`.

[26] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Journal of Machine Learning Research - Proceedings Track*, 9:297–304, 01 2010.

[27] K. He, X. Chen, S. Xie, and P. He. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[28] B. Hogan, J. Park, and R. de Lutio. Herbarium 2022 - fgvc9, 2022. Available at: `https://kaggle.com/competitions/herbarium-2022-fgvc9`.

[29] B. R. Hussein, O. A. Malik, W.-H. Ong, and J. W. F. Slik. Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review. *Ecological Informatics*, 69:101641, 2022.

[30] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics YOLOv8, Jan. 2023. Available at: `https://github.com/ultralytics/ultralytics`.

[31] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[32] F. Kleber, M. Reiter, and M. Kadic. Herbarium specimen segmentation dataset, 2024. Available at: `https://doi.org/10.5281/zenodo.11479471`.

[33] Meta AI and Papers with Code. Self-supervised image classification on imagenet, Aug 2021. Available at: `https://paperswithcode.com/sota/self-supervised-image-classification-on?metric=Top%205%20Accuracy`.

[34] A. Mora-Fallas, H. H.G. Goëau, S. Mazer, N. Love, E. Mata-Montero, P. Bonnet, and A. A.J. Joly. Accelerating the automated detection, counting and measurements of reproductive organs in herbarium collections in the era of deep learning. *Biodiversity Information Science and Standards*, 3:e37341, 2019.

[35] C. Patel, A. Patel, and D. Patel. Optical character recognition by open source ocr tool tesseract: A case study. *International Journal of Computer Applications*, 55(10):50–56, 2012.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[37] M. Rice, F. Jenkins, and T. Nartker. The fourth annual test of ocr accuracy. Technical Report Technical Report 95-03, Information Science Research Institute, 1995.

[38] L. H. Rieseberg and J. H. Willis. Plant speciation. *Science*, 317(5840):910–914, 2007.

[39] R. Smith. An overview of the tesseract ocr engine. In *Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633, 2007.

[40] J. A. Sánchez, V. Romero, A. H. Toselli, and E. Vidal. Icfhr2016 competition on handwritten text recognition on the read dataset. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 630–635, 2016.

[41] N. Tomasev, I. Bica, B. McWilliams, L. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic. Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet? *arXiv preprint arXiv:2201.05119*, 2022.

[42] N. Turland, J. Wiersema, F. Barrie, W. Greuter, D. Hawksworth, P. Herendeen, S. Knapp, W.-H. Kusber, D.-Z. Li, K. Marhold, T. May, J. Mcneill, A. Monro, J. Prado, M. Price, and G. Smith. *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017*, volume 159. Koeltz Botanical Books, 06 2018.

[43] M. Vaishnav, T. Fel, I. F. Rodríguez, and T. Serre. Conviformers: Convolutionally guided vision transformer. *arXiv preprint arXiv:2208.08900*, 2022.

[44] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748v2*, 2019.

[45] B. E. Walker, A. Tucker, and N. Nicolson. Harnessing large-scale herbarium image datasets through representation learning. *Frontiers in Plant Science*, 12, 2022.

[46] C. G. Willis, E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, G. Nelson, S. J. Mazer, N. L. Rossington, T. H. Sparks, and P. S. Soltis. Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology & Evolution*, 32(7):531–546, 2017.

[47] S. Younis, M. Schmidt, C. Weiland, S. Dressler, B. Seeger, and T. Hickler. Detection and annotation of plant organs from digitised herbarium scans using deep learning. *Biodiversity Data Journal*, 8:57090, 12 2020.

[48] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. ibot: Image bert pre-training with online tokenizer. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, 2022.

[49] Y. Zhu, T. Durand, E. Chenin, M. Pignal, P. Gallinari, and R. Vignes Lebbe. Using a deep convolutional neural network for extracting morphological traits from herbarium images. *Proceedings of TDWG*, 1:e20400, 08 2017.

Appendix

Appendix A: Visual Analysis Tool Application for botanical researchers

This thesis addresses the extraction of relevant information from these digitalised specimens, and provides the botanical researchers with a tool for the semantic analysis of the specimens. The tool allows botanical researches to explore hypothesis such as: *the size of the plant blossoms grows over the years.*. The categories they look for, in their research, are the reproductive and other organs of plants including leaves, blossoms, and fruits.

The goal of the application is to present this data to the botanical researchers in an application, so that they can explore it, using the knowledge gained from the contrastive learning model. The embeddings from the contrastive learning model are paired with the textual information, such as the dates of the collection and the geolocation of the specimens, and specimen images in an information visualization tool. More specifically, the embeddings of the model are used to calculate the K-nearest neighbours (KNN)[21] of the herbarium specimens that the researchers can interactively add to the application. The contrastive learning creates the embeddings of the images based on their features, learned in the contrastive learning process. The KNN information is then used to display the closest specimens to the selected one, which should be the most similar ones, according to the model embeddings. The users can then view the most similar specimens and use this information for further research. These specimens also have their metadata information displayed next to them, if available. This information includes the geolocation of the specimens, the family, species and the plant type according to the specimen entry, the name of the researcher or the institution that made the entry into the NHMW database, and finally the date when the specimen was collected.

The displayed specimens can also be sorted by dates, locations, and the names of the researchers or institutions that collected the specimens.

The application has two windows, one for viewing the information and another one for selecting the images from the herbarium for which the calculations are necessary. The window for the selection of the images also includes an option to fetch the ground truth information from the NHMW server, if available, and to calculate the KNN distances. There is also an option to choose the model which calculates the embeddings, the only option currently being the contrastive learning model described in the previous section of this thesis. Screenshots of the application can be seen in 1 and 2. This was done by using the a python flask backend, javascript, HTML and CSS.

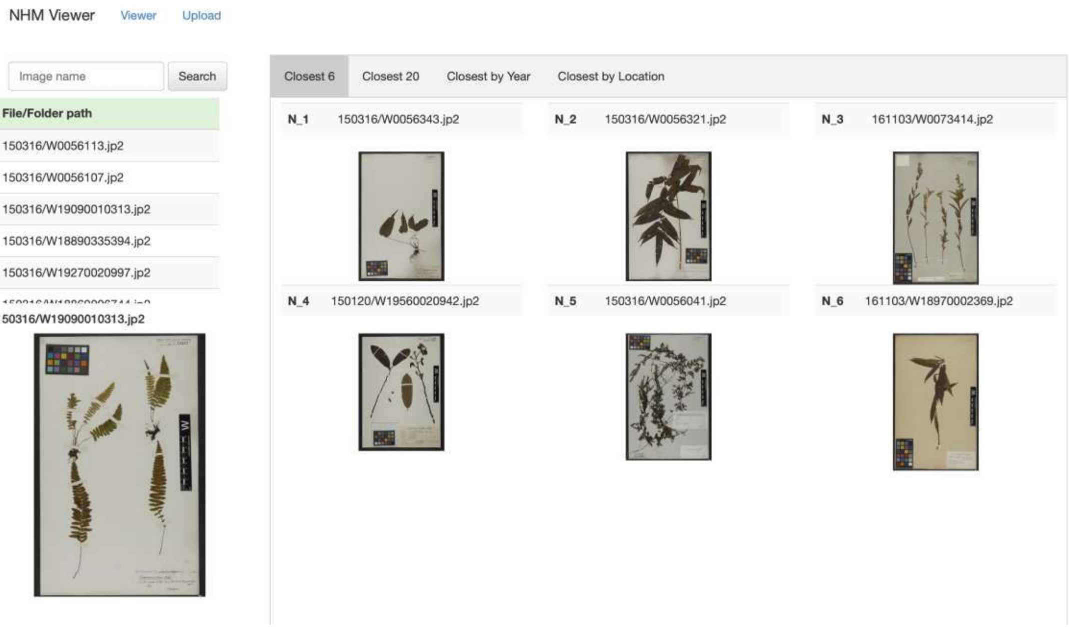

NHM Viewer Viewer Upload

Image name Search

File/Folder path
150316/W0056113.jp2
150316/W0056107.jp2
150316/W19090010313.jp2
150316/W18890335394.jp2
150316/W19270020997.jp2

150316/W19090010313.jp2

Closest 6 Closest 20 Closest by Year Closest by Location

N_1 150316/W0056343.jp2

N_2 150316/W0056321.jp2

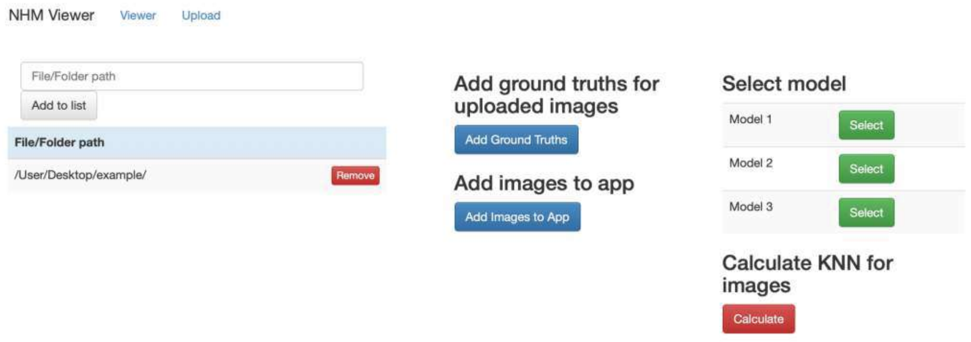
N_3 161103/W0073414.jp2

N_4 150120/W19560020942.jp2

N_5 150316/W0056041.jp2

N_6 161103/W18970002369.jp2

Figure 1: Showing the specimen similarity display window from the developed application


NHM Viewer Viewer Upload

File/Folder path

Add to list

File/Folder path	
/User/Desktop/example/	Remove

Add ground truths for uploaded images

Add Ground Truths

Add images to app

Add Images to App

Select model

Model 1	Select
Model 2	Select
Model 3	Select

Calculate KNN for images

Calculate

Figure 2: Showing the specimen upload and calculation window from the developed application