

# ML Project–

## Cross domain Plant Species Identification

Due: 11:59 pm 28/11/2025 (Friday of Week 12)

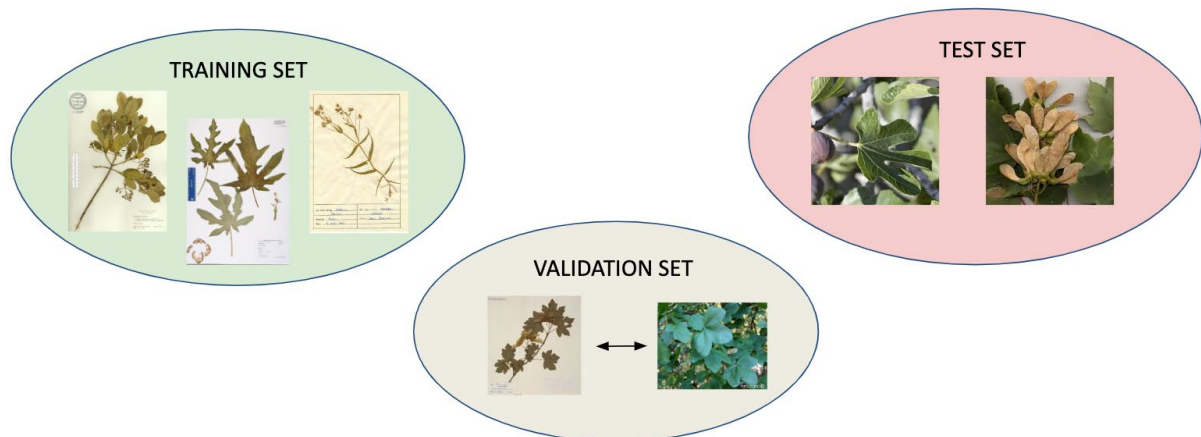
Contributes 50% of your final result

Group Assignment: Group of 4-5 students

### 1 Introduction

Deep learning has had promising success in recent years. In the context of plant species identification, especially for data-deficient species (such as those in tropical regions), field training images are still severely lacking. Therefore, identification of data-deficient plant species is a challenge. In this regard, the use of herbarium images, which are much more accessible, has been experimented with field images for the application of cross-domain plant species identification. Unlike conventional plant species identification, this cross-domain plant species identification aims to identify plant species in different domains, namely the herbarium and the field. Figure 1 shows an overview of cross-domain plant identification from herbarium images to field images.

The problem can be viewed as a cross-domain classification task. The training set will consist of herbarium sheets, *while the test set will only consist of field images*. To enable learning a correspondence between the herbarium sheet and field image domains, a *subset* of species classes has both the herbarium sheets and field images provided for validation. In this project, this subset of species classes that is treated as the validation set has been included in your training set to increase the amount of training set and to provide you with additional information to learn the correspondence between herbarium sheet and field image domains.



source: <https://www.imageclef.org/PlantCLEF2020>

**Figure 1:** An overview of cross-domain plant identification involving herbarium specimens as training dataset and field photos as test dataset. The field images are validated with the herbarium images. There is a difference in data distribution between these two domains.

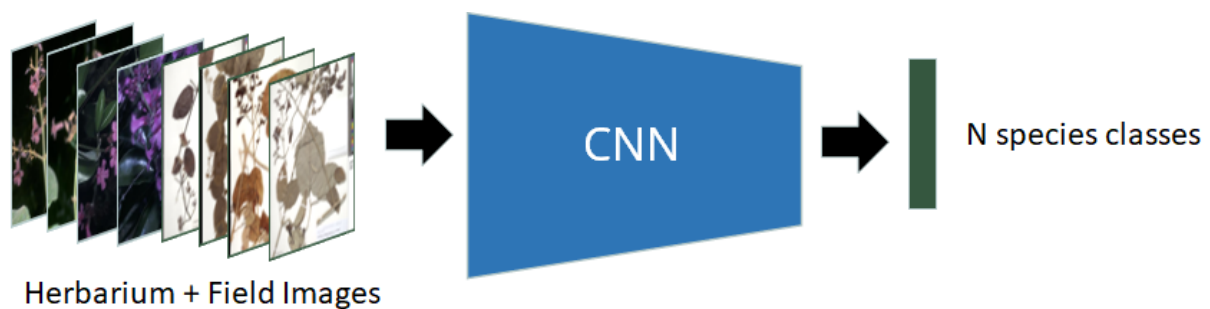
### 3. Getting Started

In this project, you are required to explore **TWO** baseline approaches and propose **A NEW** approach of your own design to improve upon the baseline methods:

*Baseline Approaches:*

#### (1) Mix-stream CNN model

As illustrated in Figure 2, this approach involves a mix-stream deep neural network architecture. You may choose any pre-trained model that you have studied previously and apply transfer learning techniques that you have learned during this experimental study. Given possible resource constraints, it is acceptable to select a lightweight pre-trained model for this task.



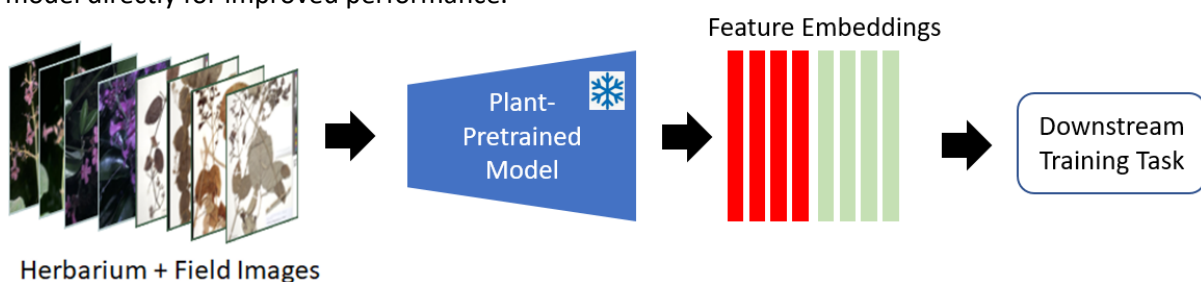
**Fig 2: The overview of mix-stream CNN model**

The CNN mix-stream must learn to match all the different image variants from different domains directly to N-species classes. Although performance may be affected by the domain shift between herbarium and field images, this is an important benchmark to test the ability of the most basic CNN approach to learn generalized features for images from different feature distributions.

#### (2) Leveraging a Plant-Pretrained Model for Cross-Domain Plant Identification

The second baseline utilizes the DINOv2 model, which has been pre-trained on plant-only images, as introduced in the 2025 PlantCLEF challenge. Specifically, you should select the version of DINOv2 that has been fine-tuned on both the backbone and the classifier head. The model can be found here: [link](#).

For this baseline, you are not expected to perform any fine-tuning on the DINOv2 model. Instead, you should use the pre-trained DINOv2 model as a feature extractor. Once the feature embeddings are generated, you may proceed to apply a similar mix-stream training approach, using traditional machine learning methods you have learned, to perform the downstream task training. However, if you have sufficient computational resources, you may optionally explore fine-tuning the DINOv2 model directly for improved performance.



**Fig 3: The overview framework of leveraging a plant-pretrained model for cross-domain plant identification**

**New Approach:**

As for the new approach, it should be based on deep learning. [Here](#) are some existing approaches where you can find ideas and references. Your model design should address the following question:

1. How to deal with classes that do not have herbarium-field pairs for the training set? Meaning some of the species classes only have the herbarium specimen but not the field samples in the training set.
2. How to deal with imbalanced classes where some have many samples in a class but others have only 1 or 2 samples?

*In this project, at a minimum, you should implement ONE new approach and compare its performance to that of the baseline approaches.*

**3.2 System evaluation**

The evaluation metric used is the top-N accuracy. The top-N accuracy is the fraction of the ground truth class being equal to any of the N highest probability classes predicted by the model. It is defined as the formula below where TP = True Positives of the N highest probability classes, TN = True Negatives, FP = False Positives, and FN = False Negatives:

$$Top-N \text{ Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

You are required to provide both the *top-1* and *top-5* accuracy results.

**4. Dataset Description**

The dataset can be downloaded here: [AML dataset](#)

It contains 100 species from [PlantCLEF 2020 Challenge](#). Herbarium and field images are provided for training, however, a subset of 40 species do not have field images. The total training images is 4,744. Meanwhile, the test set consists of 207 field images. The overview of the dataset is as follows:

Dataset	No. of images		
	Herbarium	Field	Total
Train	3,700	1,044	4,744
Test	-	207	207

File/Folder	Content
train	Herbarium and Field training images

test	Field test images
list/train.txt	List of images for training
list/test.txt	List of images for testing
list/species_list.txt	List of species and their respective classid
list/groundtruth.txt	List of ground truth for the test set
list/class_with_pairs.txt	List of species (classid) which have herbarium-field training pairs
list/class_without_pairs.txt	List of species (classid) which do not have herbarium-field training pairs

## 5. User Interface for plant species identification system

The expected input to the system is the field image and the expected output is to predict which species class it belongs to and display the relevant herbarium images it matches. The design of the graphical interface should be user-friendly and practical.

## Marking Scheme

Requirements	Mark
<b>3.1 Models Implementation (27%):</b> Implementation of the baseline approaches (12%) and the proposed new approach (15%).	27
<b>3.2 Project Report (10%):</b> A report (PDF) consisting of two sections: <i>Methodology</i> and <i>Result and Discussion</i> . Please describe your models' architecture, loss function, hyperparameters, and any other details of interest, and discussing the performance differences between them. Please limit the report to <b>8</b> pages.	10
<b>3.3 User Interface (5%)</b> The user interface requirements for the plant species identification system are listed in Section 5.	5
<b>3.4 Project Presentation (8%):</b> You are required to present your project findings to the unit lecturer and industry partners. Please <i>record your video presentation and include the link in your report</i> . The video should be <i>no longer than 8 minutes</i> .  A Q&A session will be scheduled, and the date and time will be announced on Canvas.	8
Total	<b>50</b>

**NOTE:**

- Individual marks will be proportionally adjusted based on each team member's overall contribution to the project as indicated in the '*Who did what*' declaration.
- You must also provide to [shlee@swinburne.edu.my](mailto:shlee@swinburne.edu.my) read only access to your git repository within 1 week of forming teams.

**SUBMISSION:**

- You must submit your *project report* to Canvas under ML Project submission's link by **11:59pm on 28/11/2025(Friday of Week 12)**. Standard late penalties apply - 10% for each day late, more than 5 days late is 0%.
  - Project report (max 8 pages) and link to your source code (Git-based VCS)).
- The *checklist*, *cover sheet* and '*Who did what*' declaration form, you can download them in Canvas