Data Analysis, AI and Optimization Project: To bee or not to bee

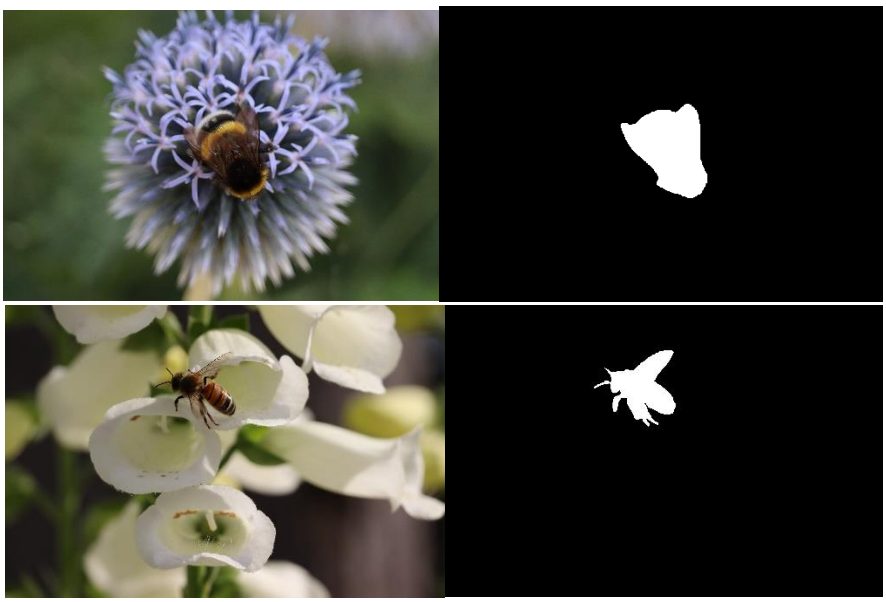
Courses : II.2413 & IG.2411, Teams of 3 to 4 students

Remark : This project is common to both II.2413 (Data Analysis) and IG.2411 (AI and Optimization) classes. If you are following only one of the two classes, you may skip the item mentioned mandatory for the other course.

I Project description

Pollinator insects such as bees and bumblebees are important for ecosystems diversity and by fertilizing flowers, they play a vital role in the global food chain system. Within this context, this project focuses on a dataset made of 347 high resolution images focused on pollinator insects with the goal of detecting and separating bees and bumblebees from other species of insects:

- Images 1 to 250 will be provided to you with a segmentation mask and an Excel File containing a two levels classification of the different insects
- Images 251 to 347 and their masks will be provided only for the last sessions, without the classification file.



Based on the first 250 images and their masks, it is expected of you to propose key features to extract and to test several Machine Learning and Deep Learning methods that you will train to identify the different types of bugs.
In the last labs, your methods will be tested on images 251 to 347, and part of you mark will depend on your algorithms ability to correctly recognize the different bugs.

II Expected deliverable

The following elements are expected as deliverables for **June 5th 2024 11:59pm**

- A PDF file report explaining the different features you tried and algorithms your tested
- A CSV file with two columns (“ID” and “bug type”) with the results of your best algorithm on images 251 to 347.

III Details on expected work

III.1 Feature Extraction [7 points]

Using the training data and masks, you are expected of extracting the following features to be used by your Machine Learning and Deep Learning Algorithms:

- Symmetry index **[IG.2411]**
- The ratio between the 2 longest orthogonal lines that can cross the bug (smallest divided by longuest) **[IG.2411]**
- The ratio of the number of pixels of bug divided by the number of pixels of the full image
- The min, max and mean values for Red, Green and Blue within the bug mask.
- The median and standard deviation for the Red, Green and Blue within the bug mask **[II.2413]**
- A least two other features of your choosing **[II.2413]**, or at least four other features **[IG.2411]**. You may use features extracted inside or outside of the bug mask.

III.2 Data visualization [5 points]

The following visualizations should be included in your report, with meaningful comments:

- Graphics of your choosing to display the repartitions between the different types of bugs (using the columns “bug type” and “species” from the Excel File)
- A PCA projection of your features into 2 dimensions
- At least 2 other projections of your features using non-linear methods **[II.2413]**

III.3 Machine Learning and Deep Learning [6+2 points]

While the specific species is provided, due to the somewhat low number of images, and the constraint to use your own features, you will have to train your Machine Learning and Deep Learning Methods to predict the “bug type” (from column 2) and not the species (from column 3). To this end, you will be expected to try Machine Learning and Deep Learning methods with the following criterions :

- 2 supervised methods that are neither deep learning nor ensemble learning
- 1 supervised ensemble learning method **[II.2413]**
- At least 2 clustering methods **[II.2413]**
- At least 1 supervised neural network using your own features **[IG.2411]**.
- A supervised method of your choosing trained over optimally auto-encoded features based on your extracted features. You will detail your auto-encoder architecture and training process **[IG.2411].**

For each algorithm, supervised or otherwise, you will compute the relevant quality indexes for the training data and comment your results.

Furthermore, during the last lab, you will have to re-apply one supervised method of your choosing to the test data: Your method will be trained on the training data, but applied and evaluated on the test data. From there, you will provide the requested CSV file from your best classification method.

Optionally, but **you may not use this method to produce your CSV**:

- Up to two supervised algorithms of your choosing trained from the raw images.
- Up to two supervised algorithms of your choosing to attempt predicting the “species” column instead of just the “bug type”.