Deep Learning Group Project
Master's Degree in Data Science 24/25

In this project, you will develop a deep-learning model to classify rare species based on their images. The dataset used in this project consists of images of rare species sourced from the Encyclopedia of Life (EOL) and curated as part of the *BioCLIP: A Vision Foundation Model for the Tree of Life* [1] study. Each image is associated with metadata, including its kingdom, phylum, and family.

Your objective is to build a model that predicts the *family* of a given species based on its image. The dataset includes a structured CSV file containing the file path of each image alongside its corresponding labels. As part of the project, you will need to create your own data splits into training, validation, and test sets. The training and validation sets will be used to develop and fine-tune your model, while the test set should be kept separate to evaluate its performance on unseen data.

Like almost every Machine Learning task, there is no "right solution" for this problem. You can create your model however you wish, **as long as you do not look at the test set.** To develop your model, feel free to apply different pre-processing techniques, test distinct combinations of hyperparameters, and make use of all the concepts covered in the course. The project should be done in Python 3. Naturally, you can (and should) use available libraries, such as Pandas, Keras, etc. Just make sure to credit everything you use in the report appropriately. Also, while the approaches you try are likely to have already been tried by someone else, **you should not plagiarise anyone's code**.

Report
In addition to your code, you should deliver a short scientific report (maximum 5 pages + bibliography, if applicable) with the following information:
Group identification of each member's student ID and name.
A clear description of your best approach (why it is the best, etc..) and the steps you followed for arriving at that model, including (when applicable):

- The applied pre-processing.
- The model/approach implemented.

-Your experimental setup:

- The list of Python libraries (for example, "*We used Keras to build the model and Keras Tuner to hyperparameter tuning*") needed to run your code and any other implementation details that can be relevant to reproduce your experiment.

-How you evaluated your approach:

- A brief description of your intermediate models (what have you tested and why it has not worked compared to your best approach, i.e. "*The standard scaler worked better instead of MinMax, etc.*.").
- An evaluation of your best approach, which should appropriate metrics. Feel free to include results for other intermediate approaches you tried.

-A summary of the results reported:

[1]: Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., ... & Su, Y. (2024). Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19412-19424).

- A brief error analysis: what are the most common errors made by your approach? Feel free to illustrate it with a confusion matrix and/or specific examples. (Maybe images of wrong classifications with the corresponding conclusion why)

-Future work:

- If you had more time, what would you have tried?

The report must be written in English.

Deliverables:

By the 30th of April, you must deliver your source code used and a report that discusses the previous points and where you explain your choices. You must deliver your work through Moodle (a Turnitin assignment will be available soon). The required files must be uploaded in a unique zip/rar file named with the group number (e.g., GROUP_1.rar). For this purpose, we have created an Excel file you must fill in with group information. Each group **MUST** consist of 5 students. There are no exceptions to this rule.

The grade of the project will be obtained by evaluating the following criteria:

- Quality of your image/data pre-processing, such as image transformations. (3 points)
- Ability to design and implement Deep Learning models. (3 points)
- Ability to utilize the pre-trained models and correctly adapt them to your task. (2 points)
- Amount and quality of innovative approaches implemented. (2 points)
- Ability to analyse the results obtained and to use Deep Learning concepts correctly. (1 point)
- Quality of the report and language adequacy. (5 points)
- The performance of your model, and the metrics used. (2 points)
- Clarity and quality of the projects code. (2 points)
- Comparative evaluations against other groups' solutions.

Beside the criteria mentioned above, your group will be compared and evaluated against other group's solutions.

The collaboration between groups and the use of AI tools to generate code is strictly forbidden. Not respecting these rules will result in the immediate reprobation of the curricular unit in both epochs.

The fact that you are working in a group does not imply that all the components of the group will receive the same grade.

[1]: Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., ... & Su, Y. (2024). Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19412-19424).