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

In recognition of **Breast Cancer Awareness Month**, in this project you will leverage deep learning to support breast cancer diagnosis through image classification. The dataset used for this project is BreaKHis and contains high-resolution microscopic images of breast tissue.
Each image is accompanied by metadata with the following features:

- **path_to_image**: The file path to the histopathological image.
- **Benign or Malignant**: A label indicating the tumor type.
- **Cancer Type**: The tumor type of each tissue, with the following labels:
  - **Adenosis**
  - **Tubular Adenoma**
  - **Phyllodes Tumor**
  - **Ductal Carcinoma**
  - **Lobular Carcinoma**
  - **Mucinous Carcinoma**
  - **Papillary Carcinoma**
  - **Fibroadenoma**
- **Magnification**: The magnification level at which the image was captured.

Project Objectives
This project has two stages:

1. **Stage 1**: Train a binary classification model to distinguish between benign and malignant images.
2. **Stage 2**: Extend the model to multi class classficiation type and predict the images by specific tumor type.

Your objective is to develop a model that can assist in the accurate diagnosis of breast tissues by classifying images in both broad and detailed categories.
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:
- 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 **15th of December**, 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)

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.