Introduction to Focus Areas in Bioinformatics – WS21/22
Lecturer: Kathleen Gallo

Project 3

- Deadline for the **REPORT**: 13.11.2021, 18:00

The Data

In this project, we want to classify data based on bio-medical (microscopy) images, namely breast cancer histopathology images.

The data is described in this paper: Spanhol, Fabio A., et al. "A dataset for breast cancer histopathological image classification." IEEE Transactions on Biomedical Engineering 63.7 (2015): 1455-1462. (www.inf.ufpr.br/lesoliveira/download/TBME-00608-2015-R2-preprint.pdf) and can be found here: https://www.kaggle.com/ambarish/breakhis.

The entire dataset contains 7909 images of breast tumor tissue with different magnifying factors. However, you can restrict yourselves to use only one magnifying factor, e.g. 400X. You can also ignore the additional cancer types and focus on the labels “benign” vs “malignant”.

The Task: Deep Learning for image classification

Your task is to develop (or: train) Deep Learning based classifiers that works directly on the images. This might get you started:

- https://towardsdatascience.com/tensorflow-2-0-create-and-train-a-vanilla-cnn-on-google-colab-c7a0ac86d61b
- https://neptune.ai/blog/how-to-use-google-colab-for-deep-learning-complete-tutorial
- https://githubmemory.com/repo/vbookshelf/Breast-Cancer-Analyzer

You should try three different Deep Learning (DL) topologies which need to be described in detail (and compared!) in the results section and to perform an evaluation using the appropriate techniques. One of the three DL topologies needs to be a CNN. One does not have to be “deep”, i.e. can have only few layers, e.g. a fully connected NN with 2 hidden layers.

The Deliverables

You need to upload all source codes and the report to your GIT repository and **send them via mail to kathleen.gallo@online.de**.

- The report should be about 1000-1500 words in length.
- The report must be delivered in PDF format (using the usual template).
- The report must contain a screenshot of the final classifier results.
- The following sections must be present (you can add more if needed):
  - Abstract
  - Scientific Background
  - Goal
  - Data
  - Results
    For each task, describe what you did (e.g. which steps you took and which methods you used-including a description) and what the results are.
  - Discussion 1: Discuss your results
  - Discussion 2: Discuss, why is this a typical project for a data-scientist? (Or why not?)
  - Appendix (Who did what?)