<u>Topic 2:</u> Properties of selected features that were derived from applying the Integrated Gradients explanation method to a Graph Convolutional Neural Network.

Supervisor: Hryhorii Chereda

Description: During the course of the development of deep learning, two areas have emerged: geometric deep learning and explainable AI. Graph Convolutional Neural Network (GCNN) is a method of geometric deep learning, that utilizes Gene Expression (GE) profiles of patients to predict their outcomes. The GE data can be structured by a protein-protein interaction (PPI) network. We use GCNN known as ChebNet (Defferard et al. *NIPS* 2016). The Layer-wise Relevance Propagation (LRP) explanation method has been applied to GCNNs and the explanation of individual classification decisions were presented via patient-specific subnetworks (see Chereda et al. *Genome Medicine* 2021). The Integrated Gradients (IG) method (Sundararajan et al. *ICML*, 2017) can also be applied to explain classification decisions. To rank and select features, these explanations can be aggregated using two different techniques. One way is to aggregate explanations over all classes for every data point (*Marcilio and Eler, SIBGRAPI, 2020*). Another way is to aggregate only the explanations that correspond to a predicted class for every data point. Later, the selected features can be analyzed according to the criteria provided by (Chereda et al. *Artificial Intelligence in Medicine* 2024). These criteria allow to estimate the following properties of selected features: stability, their impact on model's performance, and biological interpretability.

Research aims:

1. Implement the Integrated Gradients method and apply it to the Keras version of GCNN. The Keras version of GCNN is provided by the supervisor (see **Useful links** below).
2. Obtain explanations, aggregate them (*Marcilio and Eler, SIBGRAPI, 2020*), and compute the criteria for selected features as in (Chereda et al. *Artificial Intelligence in Medicine* 2024). The GCNN models will be trained by a student and the data will be provided by the supervisor.
3. Repeat the procedure above, but this time aggregate only these explanations that correspond to a predicted class for every data point.
4. Compare the results from steps one and two. The main aim is to analyze the properties of feature selection and compare between two approaches aggregating explanations.

Requirements:

- Higher than intermediate skills in python, familiarity with Keras
- Knowledge of the foundations of machine learning
- Access to NHR servers since the computations will be done there

Main papers:

H. Chereda, A. Leha, T. Beißbarth, "Stable feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation for biomarker discovery in breast cancer", *Artificial Intelligence in Medicine*, Volume 151, 2024, https://doi.org/10.1016/j.artmed.2024.102840.

Sundararajan et al. “Axiomatic Attribution for Deep Networks”, *ICML*, 2017
http://proceedings.mlr.press/v70/sundararajan17a/sundararajan17a.pdf

W. E. Marcílio and D. M. Eler, "From explanations to feature selection: assessing SHAP values as feature selection mechanism," *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Porto de Galinhas, Brazil, 2020, pp. 340-347, doi: 10.1109/SIBGRAPI51738.2020.00053.

Useful links:

The Keras implementation of GCNN can be found here:
https://gitlab.gwdg.de/UKEBpublic/graph-lrp.
A tutorial on how to apply IG to Keras models:
https://www.tensorflow.org/tutorials/interpretability/integrated_gradients.