Hierarchical Explainability

While commonly machine learning models make a single prediction, some predictive tasks require the assignment of more than one label for each instance. Furthermore, for some of these problems, the labels have a hierarchy, i.e., they are structured in the shape of trees or directed acyclic graphs (see Fig. 1 for examples of hierarchical data). For this specific class of problems, we have developed a library called HiClass [Miranda, 2023] with generic implementations of local hierarchical classification algorithms, which enables users to quickly train and evaluate hierarchical models on different application domains.

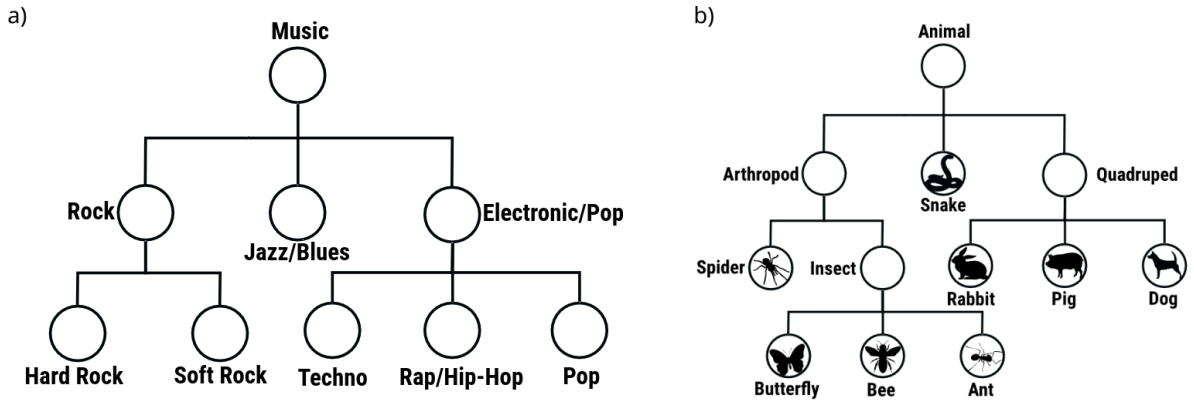

Figure 1: a) When labeling and retrieving musical information, the genre plays an important role, since having the musical genres structured in a class hierarchy simplifies how users browse and retrieve this information. b) When grouping different species, scientists typically build a hierarchical structure to distinguish organisms according to evolutionary traits.

Nowadays, many sensitive applications require insights into the workings of a machine learning model due to regulatory requirements and ethical considerations. This can be achieved by interpretable machine learning. It comes in different forms: Models can be interpretable by nature giving insights about relevant features for the prediction based, for instance, on their parameters (eg. Linear regression). On the other hand, non-interpretable models, also called black box models, are usually too complex to derive information on the features' relevance directly. Those can be analyzed with various post-hoc attribution methods, for example, SHAP values or Saliency Maps. Additionally, the scores which are assigned to the individual input features can be visualized to obtain more understandable information on what the model learned (see. Fig 2: Visualizations of attribution scores for (A) image, (B) genomic sequence, and (C) tabular data).

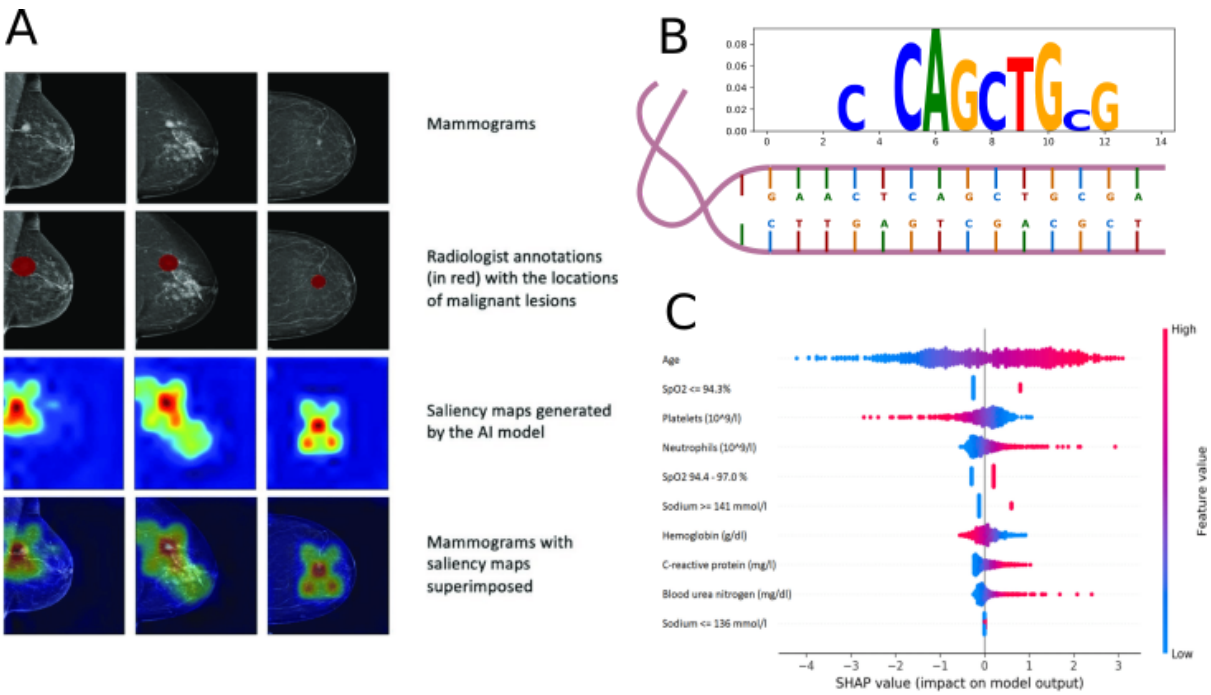

*Figure 2: Deep Learning models can be trained on various types of data so that different visualizations are needed for the interpretation of models. (A) By using saliency maps on **image data**, malignant lesions can be identified in mammograms that correlate to the annotations of radiologists. (B) Post-hoc attribution methods like DeepLIFT assign contribution scores to individual positions in a **DNA sequence** based on the impact on the model's output. Subsequences with high scores indicate biological relevance. (C) SHAP values show the impact of individual features from **tabular data** for mortality risk prediction at the start of hospitalization for SARS-CoV-2 infection. Plotting the values for all input samples can give an overview of the overall importance of each feature.*

Theoretically, there are approaches that can be transferred to hierarchical classification by either using interpretable models for the individual nodes/levels or by applying post-hoc attribution methods on each model. However, there are multiple challenges that come with that. Since there are multiple models instead of just one, it results in multiple interpretations without any hierarchical context. This leads to the following questions which we would like to address:

- How can we incorporate hierarchical information into the visualizations of the interpretability scores?
- What are the differences in interpretations between the different levels/nodes? How to describe those?
- Is it possible to infer new insights from those interpretations?
- How can we calculate and store the attribution scores efficiently without losing the hierarchical information?

The main goal of the project is to generate a prototype that solves those questions. Depending on your preferences and skills, the focus can be on one of those three tasks:

Visualization and Information

You will first read literature about hierarchical taxonomies, visualizations of taxonomies, and interpretability methods. Based on that, you will creatively develop different ways to combine hierarchical taxonomy and interpretability. This should include the comparison between

different classes/levels. You will explain changes between levels and how to represent that information so that new insights might be discovered.
Requirements: Basic data visualization and analysis skills. Participation in the seminar 'Visual Data Exploration and Story Telling - the good, the bad, and the ugly' (SS 23) is a plus.

Machine Learning
Your literature review will cover different classification methods as well as hierarchical classification. The main task here is to develop and train hierarchical models on the selected data including hyperparameter optimization. Afterward, you will interpret already interpretable models or apply post-hoc attribution methods on the black-box models. Additionally, you will examine approaches to compute and store the interpretability scores more efficiently.
Requirements: Practical experience in machine learning/deep learning in Python. Experience in explainability and interpretability is not required but is a plus.

Framework Implementation
HiClass is already implemented as an open-source Python library that is compatible with scikit-learn. It mirrors the popular API from scikit-learn to train and predict with the most common design patterns for local hierarchical classification. We would like to extend that library with a module for hierarchical explainability. First, you become familiar with the HiClass library before you design how to integrate the new explainability module. In collaboration with your colleagues from the other tasks, you will include functions (eg. I/O, visualizations) that were developed in the process.
Requirements: Practical experience in software development as well as Python. Experience with open-source Python libraries, pytest, and scikit-learn is a plus.

Why should you join this project?

Are you a creative person? Then you have the chance to investigate new ways of visualizing hierarchical model interpretations. Or are you interested in machine learning and understanding the models you train? In this case, learn about hierarchical models and apply them to various classification problems. Would you rather prefer programming tasks with user-friendly applications as your output? Extending the open-source Python library based on scikit-learn will definitely be a strong experience.
Please contact us if you don't think you meet the requirements but are interested in participating - we can adjust a task based on your preferences and skills! You are as well invited to work on multiple tasks if more grab your attention.
Look forward to engaged supervision by Prof. Dr. Bernhard Renard (bernhard.renard@hpi.de), Fabio Malcher Miranda (Fabio.MalcherMiranda@hpi.de), and Marta Lemanczyk (Marta.Lemanczyk@hpi.de) in your upcoming master project!
Please do not hesitate to contact us in case of questions.

Further Reading

Hierarchical Classification [Miranda 2023]
Interpretable Machine Learning: [Molnar 2022]

Visualizations for hierarchical data: [Kaya 2022]
Some examples of hierarchical visualizations:
https://observablehq.com/@d3/zoomable-sunburst
https://observablehq.com/@d3/zoomable-icicle
https://observablehq.com/@d3/zoomable-circle-packing

References:

Miranda, Fábio M., Niklas Köhnecke, and Bernhard Y. Renard. "Hiclass: a python library for local hierarchical classification compatible with scikit-learn." Journal of Machine Learning Research 24.29 (2023): 1-17. https://hiclass.readthedocs.io/

Molnar, C. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.)" (2022). christophm.github.io/interpretable-ml-book/

Kaya, G., Ezekannagha, C., Heider, D., & Hattab, G. "Context-Aware Phylogenetic Trees for Phylogeny-Based Taxonomy Visualization." Frontiers in Genetics 13 (2022).

References for images:

Goergen, Stacy K., Helen ML Frazer, and Sandeep Reddy. "Quality use of artificial intelligence in medical imaging: What do radiologists need to know?." Journal of Medical Imaging and Radiation Oncology 66.2 (2022): 225-232.

Murri, Rita, et al. "A machine-learning parsimonious multivariable predictive model of mortality risk in patients with Covid-19." Scientific Reports 11.1 (2021): 21136.