Project 2: Interpretability and Explainability

Interpretable and Explainable Classification for Medical Data

Teaching Assistant: Samuel Ruiperez-Campillo

Project 2 consists of three parts. The first part uses tabular, while the second uses imaging data. In the third part, you will recap your findings and answer some general questions about the methods you have seen. You will explore techniques that enable interpretable and explainable classification using shallow and deep machine-learning methods. Before starting the project, be aware of the following points:

- Submit the report and code on Moodle until **May 12th, 2025 (23:59)**
- Report
    - **Must be a PDF**.
    - The report should be **at most 4 pages** (we encourage you to use the NeurIPS paper template)
    - The report should contain key figures.
    - **You may include an 'Appendix' file of up to 5 pages for additional figures.** It should contain only titles, figures, and figure captions—**no extra text** other than captions. Figures must be clear and legible when printed on A4, including readable axes, labels, etc. Figures with unreadable fonts will not be accepted.
    - Should be **self-contained**, i.e. without references to code.
    - **Code must be handed-in too** and should follow the guidelines on the handout.
    - Underlined sections within questions specify how many points can be achieved by solving that specific subquestion.
- You will also need to hand in your code (e.g. GitHub). Please include a requirements.txt or similar for your Python environment and a README.md explaining how to run your code.
- Do not train on the test sets and only provide results that are evaluated on the test set.
- Both datasets are publicly available on Kaggle. It is allowed to use publicly available code (apart from other groups' work) but make sure to properly reference external sources.
- For computation-heavy tasks, we have arranged access to the student cluster. The datasets can be accessed on `/home/[your username]/ml4h_data/p2`.
- Please refer to the introductory tutorial slides for more information about the student cluster.

Please, make it very clear in your report where you are answering each of the questions and tasks, and, in the figure captions, indicate which part, questions and subquestions you are aiming to give an answer to.

Part 1. Heart Disease Prediction Dataset **(20 Pts)**

For Part 1, we will provide you with train and test splits from the Kaggle Heart Failure Prediction Dataset aggregated from UCI Machine Learning Repository over Moodle.

Q1: Exploratory Data Analysis **(3 Pts)**

Get familiar with the dataset by:

- *exploring the different features, their distribution, and the labels* **(Q1.1. [1 Pt]).**

Check for common pitfalls like:

- *missing or nonsensical data, unusual feature distribution, outliers, or class imbalance, and describe how to handle them* **(Q1.2. [1 Pt]).**

After having familiarized yourself with the data,

- *explain how you preprocess the dataset for the remaining tasks of part 1* **(Q1.3. [1 Pt]).**

Interpretability and explainability aim at gaining more insights about the data than just optimizing predictive performance.

Q2: Logistic Lasso Regression **(5 Pts)**

By design, linear models are interpretable due to the weights that intuitively provide feature importance values. Further, we can perform l_1 regularization to sparsify weights, allowing us to understand which features do not contribute to the outcome. For this question,

- *fit a Lasso regression model with l_1 regularization on the dataset* **(Q2.1. [1 Pt]).**

Additionally, to ensure comparability of feature coefficients,

- *describe which preprocessing steps are crucial* **(Q2.2. [1 Pt]).**

To quantify the performance of this model,

- *provide performance metrics such as f1-score or balanced accuracy* **(Q2.3. [1 Pt]).**

Lastly,

- *visualize the importance of the different features and argue how they contribute to the model's output* **(Q2.4. [1 Pt]).**

Consider the following setting: A researcher is interested in the important variables and their influence on the label. They have fitted the Logistic Lasso Regression to determine the important variables. Then, they train a Logistic Regression solely on these variables and use this model to make conclusions.

- *Elaborate why this would be a good or bad idea* **(Q2.5. [1 Pt]).**

Q3: Multi-Layer Perceptrons **(5 Pts)**

While often reaching superior performance, MLPs are generally hard to interpret, and it is not straightforward to see what is happening within these models. We thus opt for post-hoc explainability methods such as SHAP[1]. Post-hoc explainability methods typically use some procedure during inference to find the feature importance per sample.

- *Similar to Q2, implement a simple MLP, train it on the dataset, and report test set performance* **(Q3.1. [2 Pt]).**
- *Then, visualize SHAP explanations of the outputs of two positive and negative samples and feature importances of the overall model* **(Q3.2. [2 Pt]).**
- *Are feature importances consistent across different predictions and compared to overall importance values?* **(Q3.3. [1 Pt]).**

Elaborate on your findings!

Hint: There is an excellent SHAP library for python that provides many SHAP algorithms and visualizations out of the box.

Q4: Neural Additive Models[2] **(7 Pts)**

Another way to make deep models more interpretable is by careful design of the architecture. One example of such a model is the Neural Additive Model (NAM), which is an instance of the class of Generalized Additive Models[3] (GAM).

- *Read the paper about NAMs, implement the model, train it on the dataset* **(Q4.1. [3 Pt]).**

Like Q2-3, provide performance metrics on the test set.

- *Utilize the interpretability of NAMs to visualize the feature importances* **(Q4.2. [2 Pt]).**

Conceptually,

- *how does the model compare to Logistic Regression and MLPs?* **(Q4.3. [1 Pt]).**

Additionally,

- *why are NAMs more interpretable than MLPs despite being based on non-linear neural networks?* **(Q4.4. [1 Pt]).**

---

[1] Lundberg and Lee, “A Unified Approach to Interpreting Model Predictions.”

[2] Agarwal et al., “Neural Additive Models.”

[3] Hastie, “Generalized Additive Models.”

Part 2: Pneumonia Prediction Dataset **(20 Pts)**

For Part 2, download the Kaggle Dataset Chest X-Ray Images (Pneumonia)[4] – also stored in the cluster.

Q1: Exploratory Data Analysis (**4 Pts**)

Download and explore the data.

- *Explore the label distribution and qualitatively describe the data by plotting some examples for both labels* **(Q1.1. [1 Pt])**.
- *Do you see visual differences between healthy and disease samples?* **(Q1.2. [1 Pt])**
- *Describe one potential source of bias that could influence model performance* **(Q1.3. [1 Pt])**
- *How do you preprocess the data for your further analysis?* **(Q1.4. [1 Pt])**

Q2: CNN Classifier **(3 Pts)**

In Q3 and Q4, we will aim to use post-hoc explainability methods for visualizing the parts of the image that are important for the prediction of a model. To do that, first

- *design a CNN classifier for the dataset* **(Q2.1. [2 Pt])**

and then

- *report its performance on a test set* **(Q2.2. [1 Pt])**.

Q3: Integrated Gradients[5] **(5 Pts)**

Like MLPs, CNNs perform very well in tasks like classification, but lack interpretability due to their black-box nature. Like in part 1, post-hoc explainability methods are thus suitable alternatives. One class of post-hoc procedures specific to image data are methods that generate attribution maps that highlight the most important regions on which the CNN bases its predictions. For this part of the assignment,

- *implement the integrated gradients method and visualize attribution maps of five healthy and five disease test samples* **(Q3.1. [2 Pt])**

Additionally, answer the following questions. When needed, support your answers with a figure.

- *Do the maps highlight sensible regions?* **(Q3.2. [1 Pt])**
- *Are attributions consistent across samples?* **(Q3.3. [1 Pt])**
- *Does the choice of baseline input image have a big effect on the attribution maps?* **(Q3.4. [1 Pt])**

---

[4] Kermany et al., “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.”

[5] Sundararajan, Taly, and Yan, “Axiomatic Attribution for Deep Networks.”

Q4: Grad-CAM[6] **(5 Pts)**

Grad-CAM is another post-hoc method that generates attribution maps. Like in Q3,

- *implement the method and visualize attribution maps of five healthy and five disease test samples* **(Q4.1. [2 Pt])**

Additionally, answer the following questions. When needed, support your answers with a figure.

- *Do the maps highlight sensible regions?* **(Q4.2. [1 Pt])**
- *Are attributions consistent across samples?* **(Q4.3. [1 Pt])**
- *Compare your findings with Q3.* **(Q4.4. [2 Pt])**

Q5: Data Randomization Test[7] **(3 Pts)**

The paper “Sanity Checks for Saliency Maps.” introduced the data randomization test to check the trustworthiness of the saliency maps of specific methods. They propose to retrain the classifier on the train set when randomly permuting labels of all samples. Then, they compare the saliency maps on test samples for the perturbed and unperturbed classifiers. We expect the map to change if an attribution map accurately captures the relationship between instances and their labels. Conversely, if the attribution map captures another concept, e.g., acts like an edge detector independent of the label, we expect the maps to stay the same. Read the paper and

- *retrain your CNN on random training labels* **(Q5.1. [1 Pt])**

Additionally,

- perform the Data randomization Test for both Integrated Gradients and Grad-CAM **(Q5.2. [1 Pt])**

Do they pass or fail?

- Elaborate and visualize your findings **(Q5.3. [1 Pt])**

[6] Selvaraju et al., “Grad-Cam.”

[7] Adebayo et al., “Sanity Checks for Saliency Maps.”

Part 3: General Questions **(8 Pts)**

To conclude, we ask you to answer the following questions to recap and reason about the project. Please answer each question for Part 1 and Part 2 separately.

Q1: *How consistent were the different interpretable/explainable methods? Did they find similar patterns?* **(Q1.1** for Part 1 **[1 Pt]; Q1.2.** for Part 2 **[1 Pt])**

Q2: *Given the "interpretable" or "explainable" results of one of the models, how would you convince a doctor to trust them? Pick one example per part of the project.* **(Q2.1** for Part 1 **[1 Pt]; Q2.2.** for Part 2 **[1 Pt])**

Q3: *Elaborate whether the feature importances from the interpretability/explainability methods intuitively make sense to find the respective disease.* **(Q3.1** for Part 1 **[1 Pt]; Q3.2.** for Part 2 **[1 Pt])**

Q4: *If you had to deploy one of the methods in practice, which one would you choose and why?* **(Q4.1** for Part 1 **[1 Pt]; Q4.2.** for Part 2 **[1 Pt])**