ETH zürich

Project 1: Interpretable and Explainable Classification for Medical Data

Alain Ryser, Moritz Vandenhirtz

07.03.2023

Big challenge to incorporate AI in clinical practice

Most practitioners do not trust AI-based clinical tools (1)

Interpretability and Explainability of models can build transparency and trust (2)

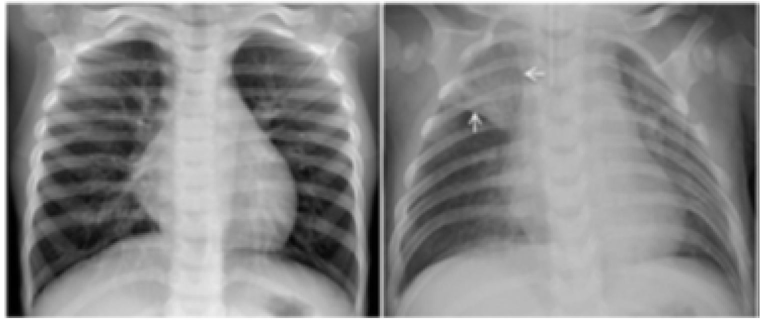
(1) https://www.fda.gov/media/143310/download

(2) He, J., Baxter, S.L., Xu, J. et al. The practical implementation of artificial intelligence technologies in medicine. Nat Med 25, 30–36 (2019)

Project 1: Interpretability and Explainability

Part 1: Tabular data for predicting coronary heart disease

Part 2: X-ray data for predicting pneumonia



Organisational

- Submit the report and code on Moodle until 18.04.2023.
- Report must be a PDF and code should follow the guidelines on the handout.
- Do not train on the test sets and only provide results that are evaluated on the test set.
- If you encounter computational difficulties due to hardware constraints, feel free to subsample or preprocess/resize the data to reduce dimensionality. Please clearly state if you apply such preprocessing, it will not impact the grading.
- 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.
- There will be a challenge task for each dataset of which you have to solve at least one to achieve the maximum grade
- The best team will be inquired to present their project for a non-cumulative bonus of 0.25 to the final grade.

General Remarks

- Please sign up your group on Moodle until 20.03.2023.
- Those that don't have a group yet can meet here after the lecture.
- As defined in the lecture, we will talk about "Interpretability" in the context of intrinsically interpretable models, while we talk about "Explainability" in the context of a black-box model that requires an auxiliary method to explain its predictions.
- Interpretability and Explainability are not an exact science. There does not necessarily have to be a single correct answer.
- A partial goal of this project is that you use the materials available to you to teach yourself the methods that are being used.

Motivation for predicting heart disease

- Cardiovascular diseases (CVDs) are by far the number 1 cause of death globally. (1)
- CVD is an umbrella term for a number of heart-related conditions.
- Of those, coronary heart disease is the most common, being responsible for 16% of the world's total deaths. (2)
- In this project, you will train ML models for early detection of potential coronary heart diseases.
- With the help of interpretability and explainability, you will gain insights into which features are important indicators.
- These insights can be helpful for understanding the disease as well as build trust of doctors toward a (semi-)automated detection of heart diseases.

(1) IHME, Global Burden of Disease (2019)

(2) https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Clinical Background (1,2,3)

- Coronary Heart Disease occurs when plaque builds up in the walls of the arteries.
- This narrows the arteries, leading to less oxygen-rich blood getting to your heart.
- Possible consequences:
  - Chest pain (=angina)
  - Heart failure
  - Blood clots that block blood flow
    - Heart muscles supplied by that artery begin to die
    - Heart attack

(1) https://www.nhlbi.nih.gov/sites/default/files/media/docs/Fact_Sheet_Know_Diff_Design.508_pdf.pdf

(2) https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142

(3) https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease

Coronary Heart Disease Prediction Dataset (1,2)

- Aggregated dataset from
  1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
  2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
  3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
  4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
- 918 Observations, 11 Features, 1 Binary response: Blood vessel diameter narrowing < or > 50%.
- Train/val - test split provided on Moodle. Please use this split for comparability and reproducibility.
- Overall goal: Develop an automated ML model for the early detection of coronary heart disease and leverage interpretability or explainability to rationalize its predictions.

(1) http://archive.ics.uci.edu/ml/datasets/Heart+Disease

(2) https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Tasks for Part 1 (detailed description of deliverables in handout)

- Task 1: Exploratory Data Analysis
  - Before training any classifier, analyse the dataset at hand in order to get an understanding of the data.
  - Based on this, preprocess the data as you see fit.
  - As first simple step to understand the relationship between response and features, compute the pairwise correlation matrix.

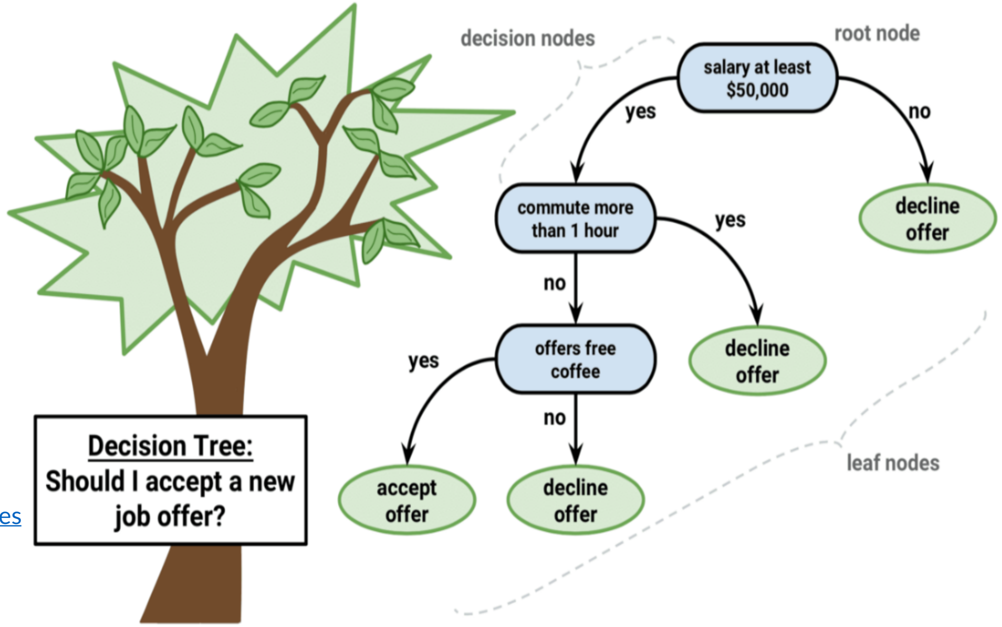
Tasks for Part 1 (detailed description of deliverables in handout)

- Task 1: Exploratory Data Analysis
- Task 2: Logistic Lasso Regression
  - Lasso combines the logistic regression with a shrinkage factor, which sets unimportant coefficients to 0.
  - This provides interpretability by differentiating between important and unimportant features.

$$\hat{\beta} = \arg\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \|\beta\|_1$$

Tasks for Part 1 (detailed description of deliverables in handout)

- Task 1: Exploratory Data Analysis
- Task 2: Logistic Lasso Regression
- Task 3: Decision Trees
  - (Inherently) interpretable
  - Feature importance by weighted reduction in impurity (1,2,3)



(1) https://towardsdatascience.com/feature-importance-in-decision-trees-e9450120b445

(2) https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn

https://stackoverflow.com/questions/49170296/scikit-learn-feature-i

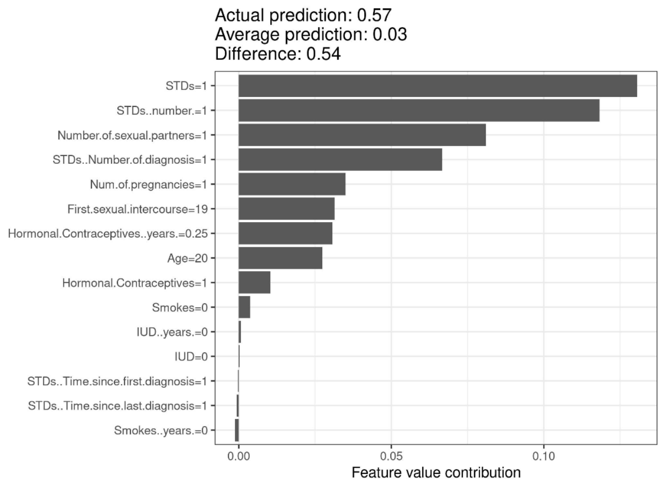
Tasks for Part 1 (detailed description of deliverables in handout)

- Task 1: Exploratory Data Analysis
- Task 2: Logistic Lasso Regression
- Task 3: Decision Trees
- Task 4: Multi-layer Perceptrons + SHAP (1)
  - MLP's considered as black-boxes.
  - SHAP is a post-hoc explanation method that computes how much each feature contributes to the final prediction for any given ML model.

(1) https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Shapley Values https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

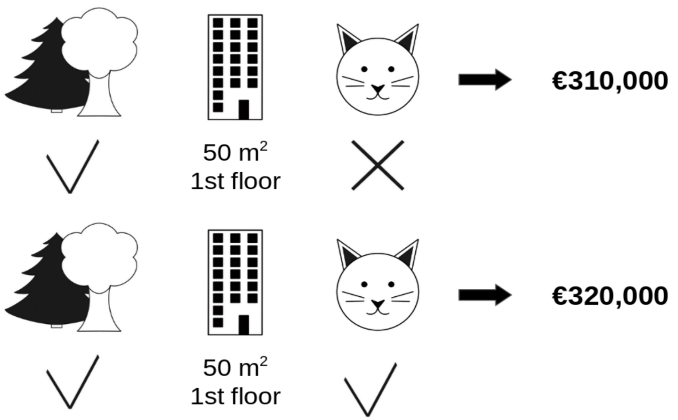
- Feature contribution metric originating from **cooperative game theory.**
- Adapted and used for **post-hoc** ML explainability purposes.
- Shapley Values are the **average marginal contribution** of each feature to the difference between the given prediction and the average prediction.
- Average prediction is computed on a **reference group.**



Shapley Values https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

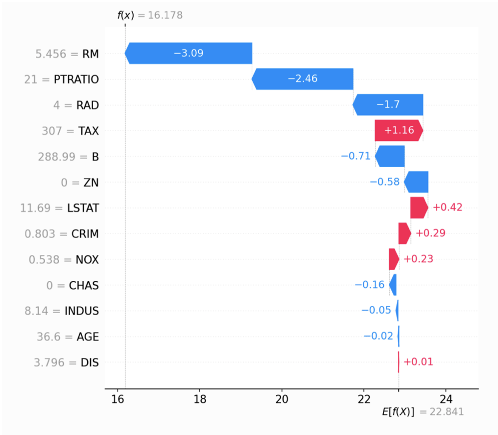
- Intuition: It's unfair to say contribution of feature to prediction is (marginal) effect of when we change its value keeping all other feature values fixed. Because the effect is also due to interactions with the other variables. Thus, one has to calculate this marginal effect for all subsets of fixed values to consider all possible interactions of features and then (weightedly) average over these subsets.

- The Interpretable ML Book gives as good explanation on how to calculate the Shapley Values.



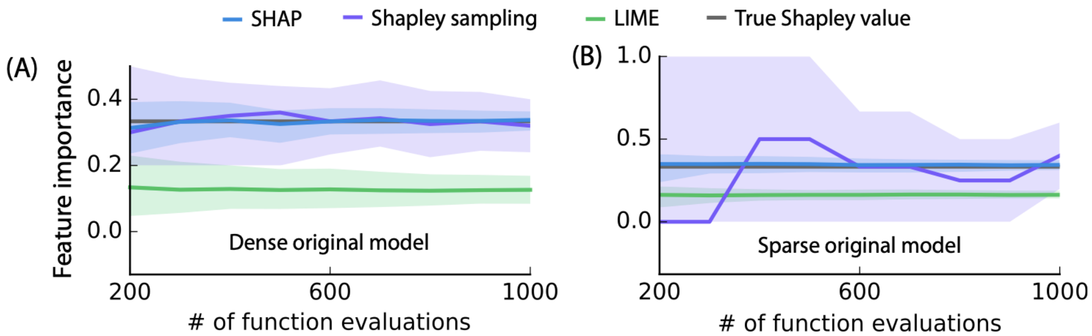
Shapley Values https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

- **Axiomatic** metric:
  - *Efficiency*: sum of values across input features add up to the difference between given and average prediction.
- **Global feature importance** by averaging absolute Shapley Values over dataset for each feature.
- Pros:
  - **Model agnostic** approach, only requires input-output pairs.
  - Information about local predictions as well as the global model.
- Cons:
  - Computation time increases **exponentially** with number of features.
  - Requires **approximations** in most real-world datasets.
  - Understanding of Shapley Values is not straightforward.



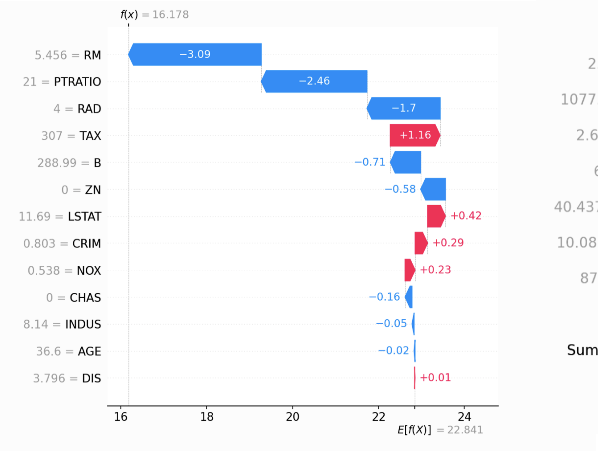
SHapley Additive exPlanations (SHAP)

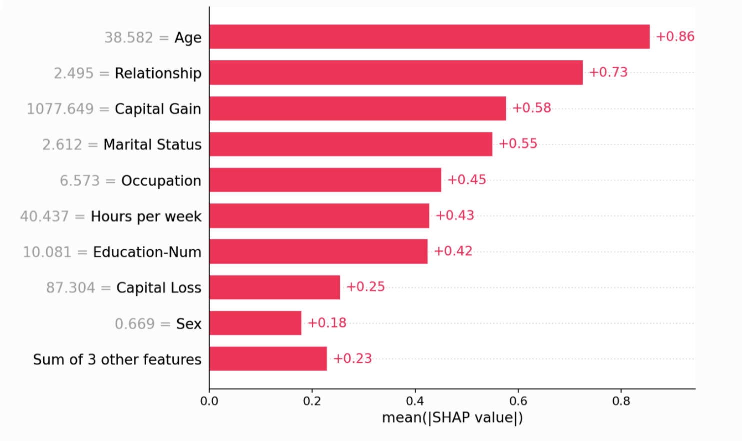
- **SHapley Additive exPlanations** (SHAP) combines sampling with other explainable approaches (i.e. Local Surrogate Models - LIME, DeepLift):
  - Increased **computational efficiency.**
  - **Improved approximation** of true Shapley Values.
- Various approximation methods :
  - Kernel SHAP: model-agnostic
  - Linear SHAP: independence assumption
  - Deep SHAP: deep networks

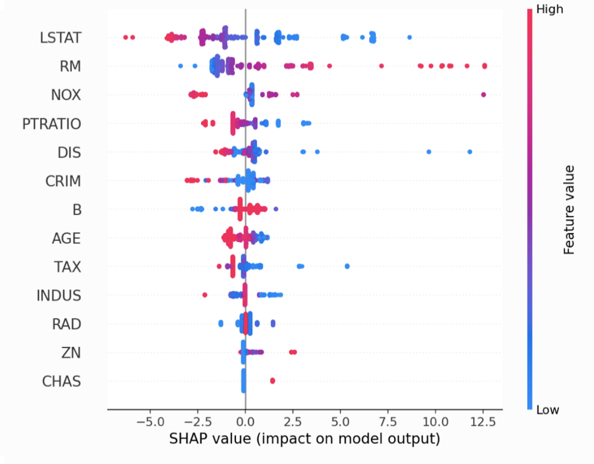


SHAP Python library https://shap.readthedocs.io/en/

- Provides automatic estimates of Shapley values for wide range of ML/DL models types. (1)
- Provides visualisations of Shapley values for explainability purposes. (2)





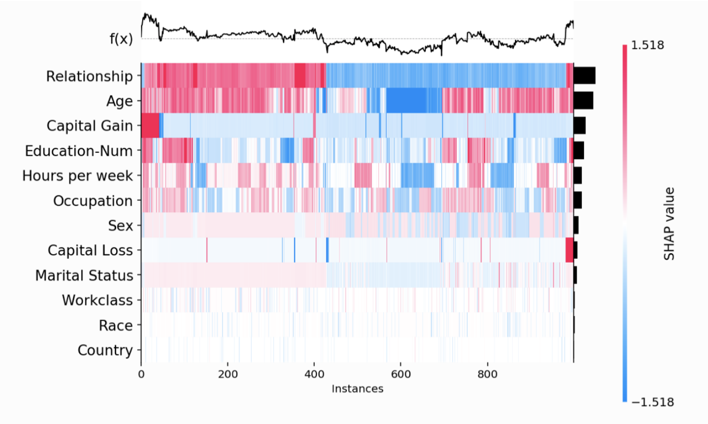


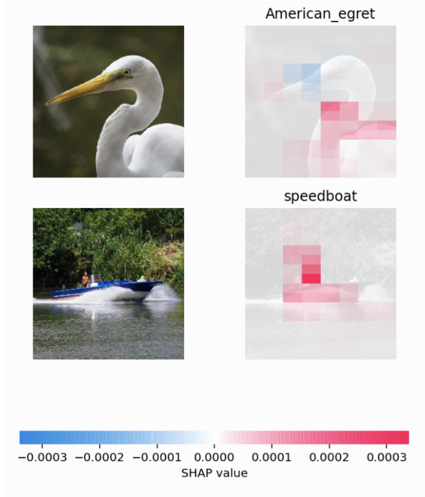
(1) https://shap.readthedocs.io/en/latest/api.html#explainers

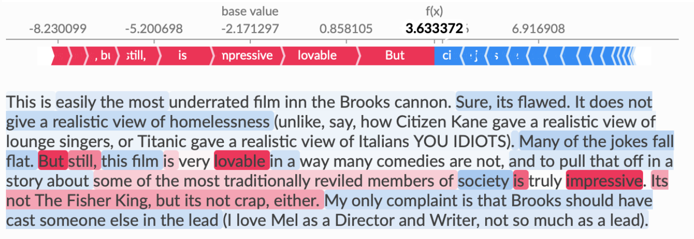
(2) https://shap.readthedocs.io/en/latest/api.html#plots

SHAP Python library https://shap.readthedocs.io/en/

- Provides **automatic estimates** of Shapley values for wide range of ML/DL models types. (1)
- Provides **visualisations** of Shapley values for explainability purposes. (2)



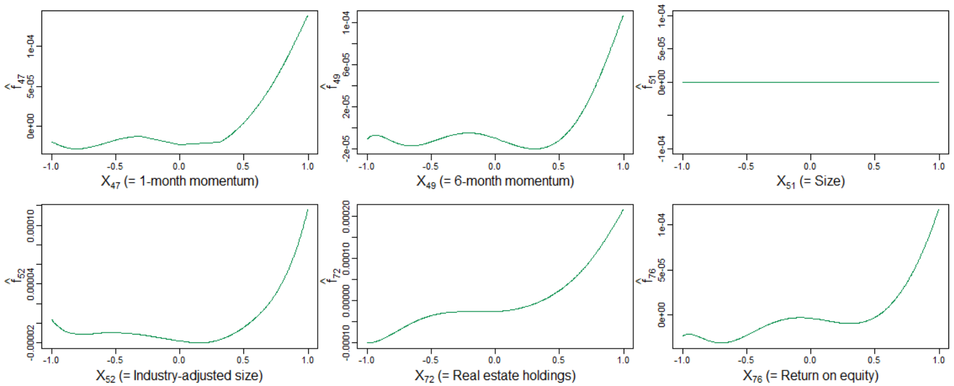




(1) https://shap.readthedocs.io/en/latest/api.html#explainers

(2) https://shap.readthedocs.io/en/latest/api.html#plots

Tasks for Part 1 (detailed description of deliverables in handout)



- Task 1: Exploratory Data Analysis
- Task 2: Logistic Lasso Regression
- Task 3: Decision Trees
- Task 4: Multi-layer Perceptrons + SHAP
- **Challenge**: Neural Additive Models (1)
  - Instance of Generalized Additive Models(2) with NN as functions

$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

  - Where we choose
  - Interpretable

(1) https://arxiv.org/pdf/2004.13912.pdf

(2) Hastie, T. J.; Tibshirani, R. J. (1990). Generalized Additive Models. Chapman & Hall/CRC.

Motivation for predicting pneumonia

- For US adults, pneumonia is the most common cause of hospital admissions other than women giving birth (in 2011). (1)
- Pneumonia is the worldwide leading cause of death for children under 5 (in 2017). (2)
- Number of radiologists is growing slower than number of images they need to analyse. (3)
- In this project, you will train ML models to detect pneumonia from X-rays.
- With the help of interpretability and explainability, you will gain insights in whether your model uses the medically revelant parts of the image for its prediction or if it uses some biases / shortcuts.
- These insights builds trust in radiologists towards a (semi-)automated detection of pneumonia.

(1) https://academic.oup.com/ehjopen/article/1/1/oeab001/6294753

(2) https://ourworldindata.org/pneumonia

(3) https://www.diagnosticimaging.com/view/are-we-prepared-for-a-looming-radiologist-shortage-

Medical Imaging Techniques

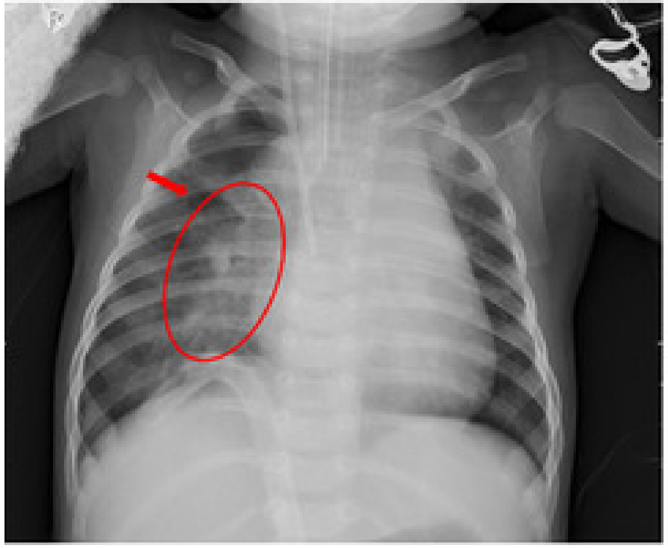
Medical Imaging Techniques are at the heart of our medical systems:

- Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Echographies , X-rays, …
- 1.18M CT exams & 1.06M MRI exams per year in Switzerland in 2019(1)
- Prevalence is increasing in Switzerland (+25% of scanners in CH over the past 5 years)(1)
- … increased workload for practitioners
- Interest from investment funds into Medical Imaging AI Companies (estimated 600M$ in 2020)(2)

(1) https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/press-releases.assetdetail.16584130.html
(2) https://www.signifyresearch.net/medical-imaging/vc-funding-for-medical-imaging-ai-companies-tops-2-6-billion

Clinical Background

- Pneumonia is an infection that inflames the air sacs (alveoli) in one or both lungs.
- Typical symptoms: Chest pain when breathing, cough, fever, shortness of breath.
- To diagnose pneumonia, the infected tissue will show denser areas and therefore appear as white spots in the darker background of the lungs.



(1) https://healthmatch.io/pneumonia/pneumonia-chest-xray

(2) https://www.lung.org/lung-health-diseases/lung-disease-lookup/pneumonia/symptoms-and-diagnosis

(3) https://www.mdpi.com/2227-7390/8/9/1423

Chest X-Ray Dataset (1,2)

- Dataset of 1-5 year old children from Guangzhou Women and Children's Medical Center.
- 5,863 X-ray images (JPEG) annotated with pneumonia/no pneumonia.
- Train - val - test split provided directly on kaggle. Feel free to rearrange train/val but please keep the testset as-is for comparability and reproducibility.
- Overall goal: Develop an automated ML model for the detection of pneumonia and leverage interpretability or explainability to investigate whether it uses the medically relevant features for its prediction.

(1) Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." cell 172.5 (2018): 1122-1131.

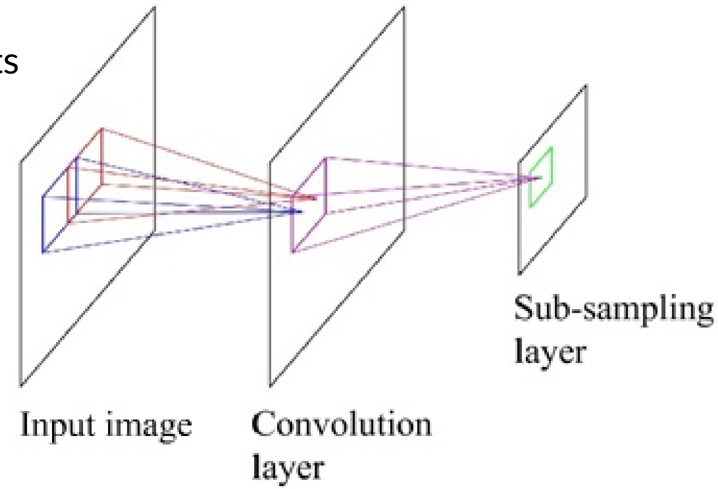
(2) https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

Tasks for Part 2 (detailed description of deliverables in handout)

- Task 1: Exploratory Data Analysis
  - Before training any classifier, analyze the dataset at hand in order to get an understanding of the data.
  - Do you see visual differences between healthy and disease patients?
  - Based on your findings, preprocess the data as you see fit.

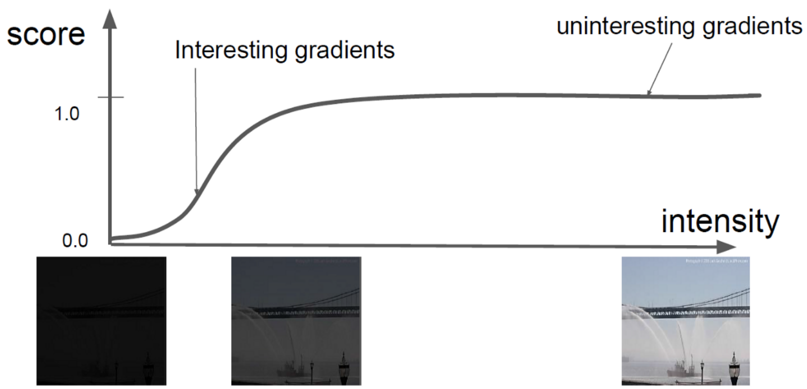
Tasks for Part 2 (detailed description of deliverables in handout)

- Task 1: Exploratory Data Analysis
- Task 2: CNN Classifier
  - For Task 3 & 4, we want use post-hoc explainability methods, which visualize the parts of the image a classifier utilizes for its predictions.
  - Thus, first design a small CNN classifier on which we can lateron apply these methods.



Tasks for Part 2 (detailed description of deliverables in handout)

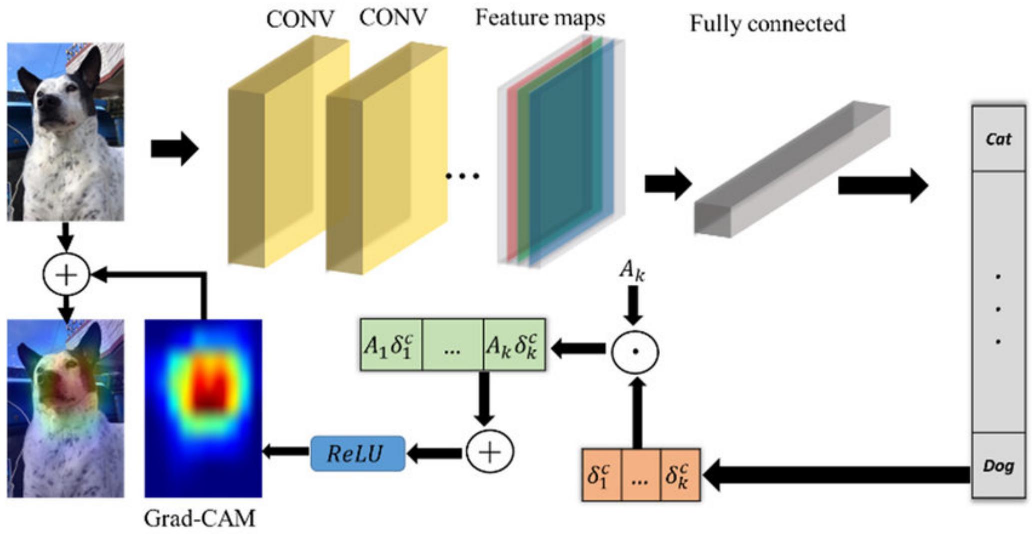
- Task 1: Exploratory Data Analysis
- Task 2: CNN Classifier
- Task 3: Integrated Gradients (1)
  - Gradient-based post-hoc methods take inspiration from the idea that the gradient of the prediction loss with respect to the input pixels indicates how important they are.
  - Gradients at image might not accurately describe importance of pixels for prediction.
  - Integrated Gradients sums up gradients over the path from a baseline (usually black image) to the image.



(1) Sundararajan, Taly, and Yan, "Axiomatic Attribution for Deep Networks."

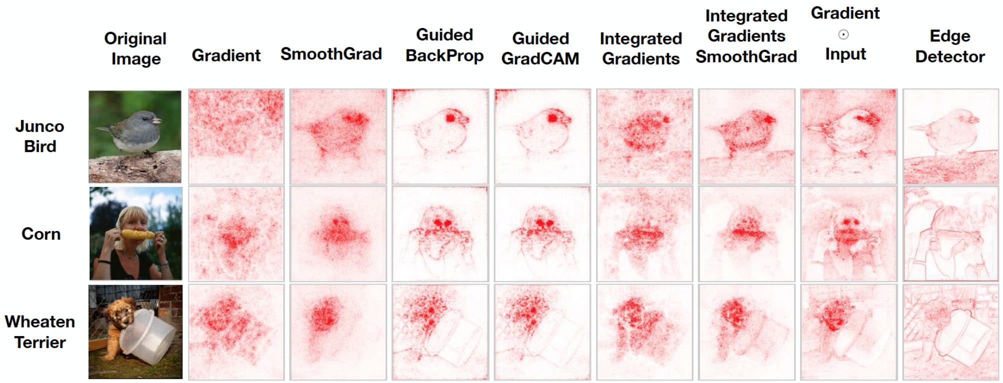
Tasks for Part 2 (detailed description of deliverables in handout)

- Task 1: Exploratory Data Analysis
- Task 2: CNN Classifier
- Task 3: Integrated Gradients
- Task 4: Grad-CAM (1)
  - Grad-CAM uses activation maps of last convolutional layer from forward pass, weighs them by gradients and upsamples them to input image.



(1) Selvaraju et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization"

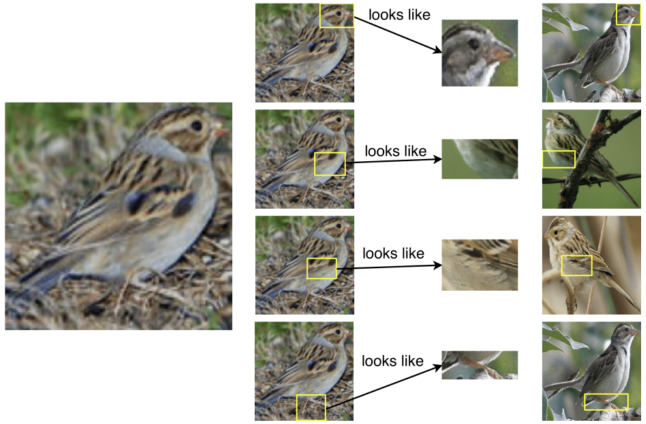
Tasks for Part 2 (detailed description of deliverables in handout)



- Task 1: Exploratory Data Analysis
- Task 2: CNN Classifier
- Task 3: Integrated Gradients
- Task 4: Grad-CAM
- Task 5: Data Randomization Test (1)
  - Show that many gradient-based post-hoc explainability methods work as edge detectors.
  - When randomly permuting labels of images, the resulting “important pixels” for classification should be random and non-sensical.
  - Instead, many methods still visualize similar important pixels as before randomization.
  - Effectively showing that methods are independent of model and classification task at hand and, thus, a bad explainer of the model.
  - A premise of interpretable models is that this would not happen for them

(1) Adebayo et al., “Sanity Checks for Saliency Maps.”

Tasks for Part 2 (detailed description of deliverables in handout)



- Task 1: Exploratory Data Analysis
- Task 2: CNN Classifier
- Task 3: Integrated Gradients
- Task 4: Grad-CAM
- Task 5: Data Randomization Test
- Challenge: Prototype Learning (1)
  - Prototypes can be either determined beforehand (1) or jointly with classifier. (2)
  - Given a set of prototypes, one can use the k-nearest neighbor algorithm for prediction.
  - In this project we will focus on the simple case of first learning prototypes using maximum mean discrepancy and then predicting the label given these prototypes.
  - Interpretable

(1) Kim, Khanna, and Koyejo, "Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability."

(2) Chen et al., "This Look Like That: Deep Learning for Interpretable Image Recognition."

Part 3

- General questions allowing you to reflect on your learnings during the project

Supplementary Material

https://christophm.github.io/interpretable-ml-book/