MLHC-Project 1

Tim Flück, Adriana Brenzikofer, Daniel Ferster

April 2023

Part 1: Heart Disease Prediction Dataset

Q1: Exploratory Data Analysis

The heart disease dataset has 12 columns: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST Slope and HeartDisease. 5 columns were categorical datatypes, so we represented each category in a column by a value so it is easier to work with for the model. We saw there was no missing data, but some data was probably default 0 and not filled in. Resting BP would probably never be 0 and Cholesterol neither. So we replaced the data containing 0 in the columns Cholesterol and BP with their mean.

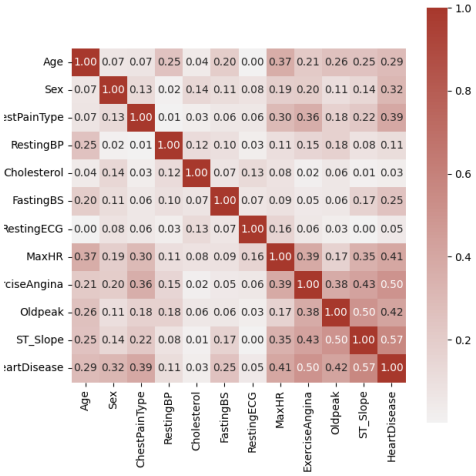

Figure 1: Correlation Matrix

Q2: Logistic Lasso Regression

A preprocessing step that is crucial to ensure comparability of feature coefficients is to turn each Feature that is of categorical type to a numerical, which we have done previously. The coefficients of a feature signify the importance of each feature for the Logistic Lasso regregression. The dropped columns are: Age, cholesterol, Resting ECG. After having removed these columns, the balanced

accuracy score remains the same at around 0.831. We used the new training and test set for other models but the original one with more columns had better accuracy scores, so we kept the old dataset.

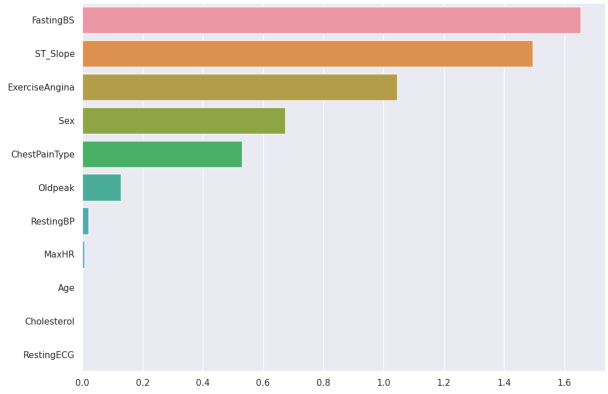

Figure 2: Logistic Lasso Regression Importances

Q3: Decision Trees

Classification performance on the test set is 1.

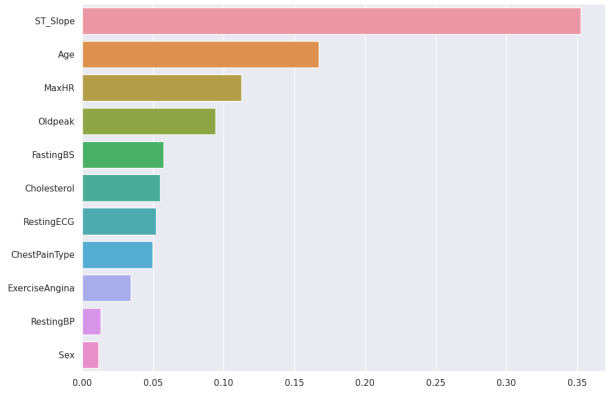

Figure 3: Decision Tree Importances

Q4: Multi-Layer perceptrons

The optimization procedure we used is Adam, the activaton function used is relu. We used sklearns MLP with the default values. Feature importances are not consistent. The accuracy score on the

test set is around 0.929

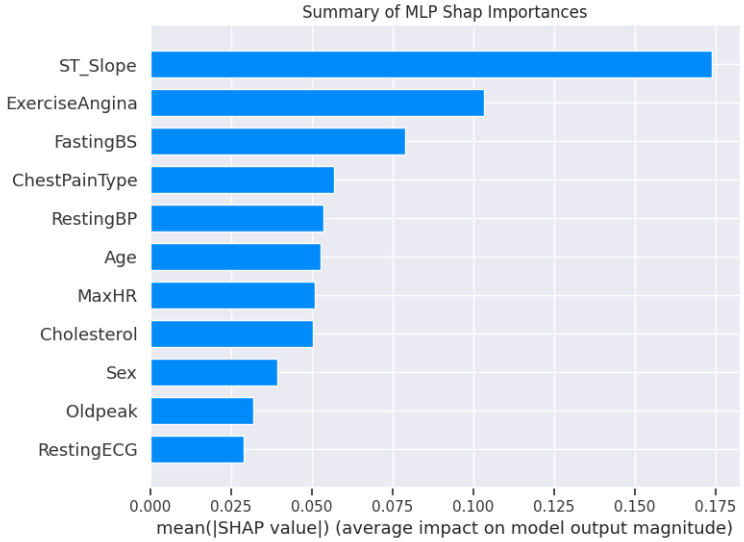

Figure 4: MLP Shap Importances Summary

Part 2: Pneumonia Prediction Dataset

Q1: Exploratory Data Analysis

When we plot the first 5 X-ray images of the training set for healthy patients (NORMAL in the left column) and patients with Pneumonia (PNEUMONIA in the right column) we can see that the X-ray images of the patients with Pneumonia show a more cloudy texture in the area of the Lung.

We can also see that the pictures have different dimensions. That is an image property that has to be standardized before working with the images.

Another difference is the font difference of the R between pneumonia and normal images which could be sources of bias and prevent learning from the Xray images for the model. Therefore the images need to be cropped so they don't contain the R.

We preprocessed the images through the keras preprocessing class and resized all images to a size of 256x256, normalized pixel values between 0 an 1, and added some data augemntation methods to the training data.

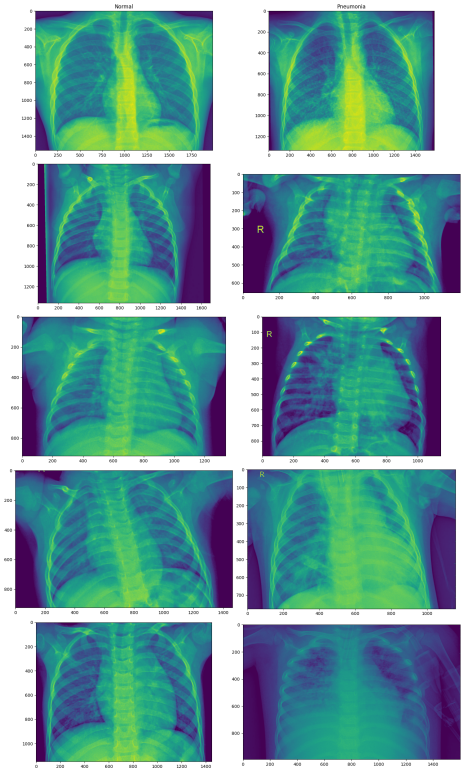

Figure 5: Normal vs Pneumonia images

When we plot the number of samples in the dataset we can see, that the training and the test data is very unbalanced. There are more Pneumonia samples than Healthy samples in both, the train and test dataset. This is another thing we have to consider during the preprocessing steps. Also the validation dataset is very small. We did not change the data distribution because of time restraints but we did add a class-weight balancer to the model during fitting to counteract the unbalanced training data.

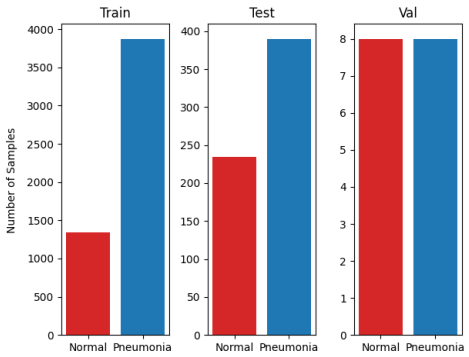

Figure 6: Datasize of Train Test and Validation set

Q2: CNN Classifier

Our CNN classifier model has accuracy 0.69.

We built the CNN according to a documentation of Hardik Deshmukh on towardsdatascience.com (https://towardsdatascience.com/medical-x-ray-%EF%B8%8F-image-classification-using-convolutional-neural-network-9a6d33b1c2a)

We used the keras Sequential class to build a sequenial CNN with 4 convolution layers (8,16,32,32 filter values) with 4 2d-maxpooling layers in between and 3 fully-connected dense functions. For the activation functions we chose "relu" and for the last dense layer we chose the "sigmoid" activation function because it is a classification model. The model consists of 826,401 trainable parameters. We tried different models but all other models had a very bad performance so we decided to use this model for the predictions. We compiled the model with the adam optimizer and the binary crossentropy loss. For the metrics we chose accuracy and AUC.

Q3: Integrated Gradients

The maps don't seem to highlight sensible regions, mostly edges. Attributions don't seem to be consistent across samples, but maybe with a better model it would.

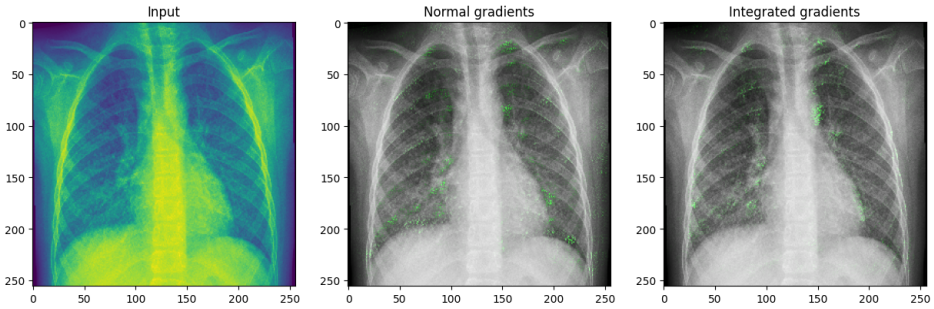

Figure 7: Integrated Gradients

Q4: Grad-CAM

For implementation of the Grad-CAM algorithm we followed the official documentation of Keras and applied it to our dataset.

The Grad-CAM highlights the lung area, sometimes with a focus on certain features. When the letter "R" was visible in the original image, the network focused on the letter for classification and ignored the rest of the image. In that way it does highlight sensible areas, especially compared to the randomised label dataset. Still, we believe the results would be clearer, i.e. more distinctive if we were to use a better model with a higher prediction accuracy.

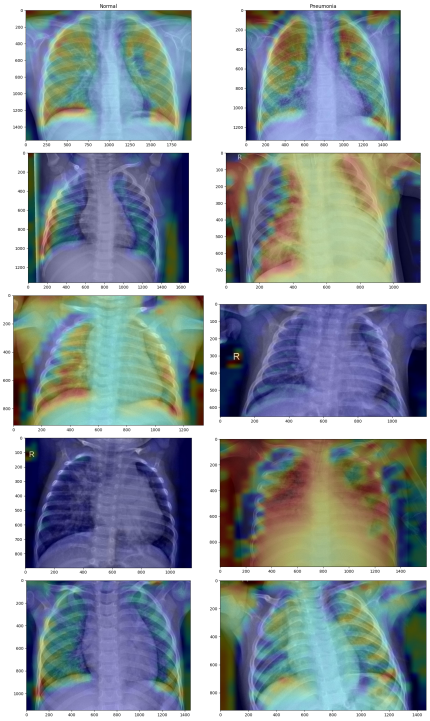

Figure 8: GradCAM

Q5: Data Randomization Test

According to the test the Integrated Gradients failed, since the result was very similar in both our model and the randomized label dataset. The Grad-CAM passed since there was a clear difference in performance between both datasets.

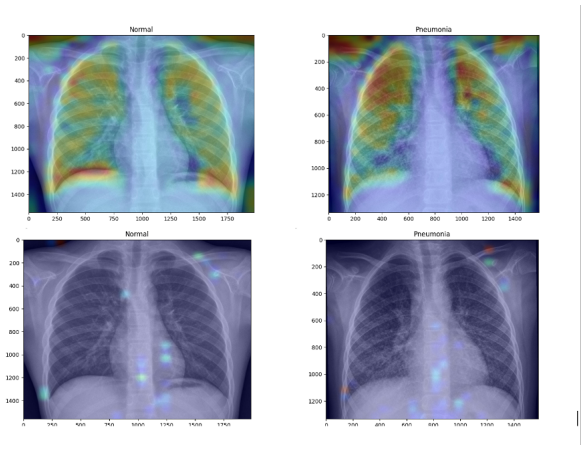

Figure 9: GradCAM on normal dataset(upper) vs randomised labels dataset(lower)

Part 3: General Questions

Q1: How consistent were the different interpretable/explainable methods? Did they find similar patterns?

Part 1: An interpretable method was the Logistic Lasso Regression. The three most important features were FastingBS, STSlope and ExerciseAngina. The explainable methods were Decision Trees and MLPs. Decision Trees most important features: STSlope, Age, MaxHR. MLP most important features overall: ExerciseAngina, FastingBS, STSlope.

For MLP the features were very inconsistent, depending on the patient you could have different weighting of the features to produce a prediction. Overall, STSlope seems the most consistent indicator for all these models for Heart Disease. Lasso Regression and MLP found similar feature importances.

Part2:

The GradCam wasn't able to recognize distinct Pneumonia streaks within the images, but it did recognize general features of the lung. That means it didn't just throw darts in the dark, the way we have observed it when applying a randomized label dataset to it. A problem we encountered was that the GradCam focused strongly on the letter 'R' which was present in some of the images, indicating the right side of the image. In those cases the rest of the image was mostly ignored, which shows that the trained network weighted the letter strongly for the prediction. The integrated gradients have shown themselves to be unreliable when compared to the randomised label set. With our model they are basically not more than sophisticated edge detectors.

Q2: Given the "interpretable" or "explainable" results of one of the models, how would you explain and present them to a broad audience? Pick one example per part of the project.

Part 1:

Logistic Lasso Regression is used to predict a binary outcome, whether or not someone has heart disease or not. The model is called Lasso because of the L1 regularization it uses, which shrinks the coefficients of some of the input variables to zero which means this feature has very low, practically no importance for the prediction for Lasso Regression. Doing so increases interpretability of the model.

Part2:

The GradCam is a visualisation method that shows the user of the ML algorhithm where it is looking. So, if you want to know on which basis it differentiates cats from dogs you would see how the algo 'looks' at those animals' snouts and ears and judges accordingly. Basically, it 'looks' at the important features of the animal or the object it analyzes. In a way, it's similar to how we humans analyze these animals. We subconsciously focus on distinctive facial features or the shape of its body, not a random blobby patch on its fur.

Q3: Did you encounter a tradeoff between accuracy and interpretability/explainability?

Part 1:

Logistic Lasso Regression shrinks some of the coefficients to zero. Removing those features can increase interpretability and we saw that for Linear Regression the accuracy stayed the same using only the important features. We did not encounter a tradeoff between accuracy and interpretability with the lasso regression because the interpretability of the features are part of the model because it automatically gives the weights of the different features. Even when leaving out the features that

had a weight of almost 0, the performance did not change. We encountered a tradeoff when we tried to use the new dataset with only the important features from the Lasso Regression on the other models. Then the accuracy was much worse! Meaning that those missing features were very helpful for these explainable models to make predictions.

Part 2:

We believe there was no tradeoff between accuracy and explainability, since both methods were post-hoc and did not influence the model, thus also didn't have an effect on its accuracy. Of course, both methods improve interpretability. In our case it was only the GradCam, but theoretically the integrated gradients should be also suited for this job.

Q4: Do your findings from the interpretability/explainability methods align with the current medical knowledge about these diseases? You may take inspiration from the references of the project presentation.

Part 1:

Current medical knowledge indicates that ST-segment elevation indicates a blockage of a coronary artery, which could lead to heart disease. This aligns with how the models laid high importance on STSlope. Surprising was to see that the MLP model did not lay a high importance to the feature Age in general.

Part 2: Pneumonia is observable with the eye on an x-ray scan by an increased opacity of lung regions. Of course, this observation method is not totally reliable. Due to this it is interesting to implement an ML algorhithm that can have a higher reliability than the human eye. So, in that way the ML method aligns with current medical knowledge, at least with regard to the analysis of x-ray images.

Q5: If you had to deploy one of the methods in practice, which one would you choose and why?

Part 1:

In Part 1 it would be MLP. This is due to the best accuracy of the model compared to the others. Of course, interpretability is the highest in linear regression, but in a sense we valued accuracy of the model higher than its interpretability. Looking at both properties together, the MLP's were the clear winner compared to linear regression and decision trees.

Part 2:

With regards to image analysis we would clearly prefer the GradCam. This we analyzed from Q5 of part 2 in which we compared the models performance to a randomized label dataset. The results of the integrated gradients were pretty similar in both our trained model and the randomized label dataset, indicating that it was functioning as an edge detector. The same test with the GradCam showed that it actually loses performance in the randomized label dataset, such that it basically has the accuracy of a coinflip. This shows that the GradCam actually shows what focus points our model has on the input images. Improvements could be made by removing the letter R from the images, because our model was confused by it, i.e. it weighted it strongly, as seen with the GradCam. Furthermore, we believe if we would've used a large scale image analysis model like VGG16, that resulted in higher accuracy in predicting pneumonia, and applied GradCam via this model, this would've resulted in more clearer distinction of pneumonia streaks in the lung.