Part 3

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

Part 1

We notice, the feature importances of logistic regression, the decision tree and the 2 layer NN show some differences.

- For example the decision tree has only a few important features, with more than half of them being very small.
- On the contrary in the case of logistic regression, we notice a more even spread of feature importances.
- For the neural additive models: we get a totally different picture with only 4 features having non-zero importance.

Overall we observed that, there is some overlap of features of high importances, but they do not completely align across all methods used.

Part 2

The method of integrated gradients was not helpful, which stands in contrast to GradCAM. Because the latter highlighted sensible regions and did produce different visualizations for healthy and disease samples. Whereas in the case of integrated gradients, we could not make out any distinction in the visualization between the two different classes. So we conclude that these two methods were not consistent.

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: Given some numerical tabular data, logistic regression derives coefficients for each feature of the data sample. The magnitudes of these coefficents are explainable by themself since they are only multiplied with the features and then summed up to get the output. Thus if the input features are of comparable size (normalization), then the coefficients can explain how important individual features are to arrive at a decision.

Part 2: GradCAM is a method used for models with images as inputs. It highlights regions of interest in the images, which the model might use for its decision. For example if a model is used to discriminate whether an image shows a dog or a cat, in the case of an image of a dog GradCAM is likely highlight the face and ears of the dog.

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

In Part 1, for the Neural Additive Models (NAMs) there was such a tradeoff. A simple 2 layer neural network performed better than the NAM in F1 and Balanced Accuracy score. But on the other hand the NAM more interpretable in terms of feature importances. Because it only had 4 non-zero absolute feature importances across all samples and for the 2 layer neural network the picture was not as clear, i.e. each feature had at least some contribution to the output.
Similiary both logistic regression and the decision performed worse than the 2 layer neural network as well. But one could argue that they are more interpretable given their simpler structure and already built-in feature importances with the magnitude of the log coefficients and the gini impurity.

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.

Heart Disease

In our case, the different models for heart disease were not that congruent in terms of feature importance. Nevertheless, one could observe a trend of features which were quite common among the different models such as: Up, ASY and SEX. Comparing these to the current medical knowledge, they roughly align with that since for example an up slope in case of ST depression during an electrocardiogram may indicate cardiac ischemia in combination with different cardiac symptoms. Furthermore, sex is a predictive feature as well, due to the fact that males are more oftenly affected by heart disease than women. Additionally, asymptomatic chestpain is a characteristic issue of heart related disease as well.

However, in medical literature the above named features do not count to the most indicative risk factors known. For example, cholesterol, resting blood pressure and diabetes are more often related to cardiac disease than the above named features. Especially that cholesterol was not even in the top 3 feature importances across our different models is suprising, since there is a proven causal relationship between heart disease and plaque formation/narrowing of the arteries due to high cholesterol levels.

As a result, our implemented models did not have the highest feature importances related to the most indicative features/ risk factors acknowledged by medicine but at least chose features which can be predictive for heart disease as well.

Pneunomia

Looking at the GradCAM, one can clearly conclude that it highlighted medically relevant regions for classification which one can especially observe in the pneumonia positive sampples. This is due to the fact that in majority parts of the left thorax get a lot of attention because the heart tends to cover major parts of the right thorax which is thereby not useful for a successful detection of pneumonia anymore. On the contrary, in pneumonia negative cases one can rather observe evenly distributed attention profiles which seem to be rather random.

Giving attention to the left thorax makes medically sense as well due to the fact that the detection of liquid in regions with a lot of contrast is rather easy compared to the right thorax which is partially covered by the heart (no/little contrast). Based on the attention profile of our trained model, we can therefore conclude that medically relevant features were extracted and as a result aligns with current medical knowledge.

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

We would choose to deploy GradCAM. It is easily applied to any model with convolutional layers and the computation is very low, i.e. it is comparable to a single backward pass during training. The resulting heatmaps are often very intuitive and can be interpreted nicely. They can be used to reason, why a model might do wrong predictions and can thus also help to properly identify issues such as bad performance and overfitting.