Project 1 - Machine Learning for Health Care

Interpretable and Explainable Classification for Medical Data

Kári Rögnvaldsson (kroegnvaldss@student.ethz.ch) Group **26**
Jacob Hunecke (jhunecke@student.ethz.ch) April 13, 2024

1 Part 1: Heart Disease Prediction Dataset

1.1 Question 1: Exploratory Data Analysis

1.1.1 Introduction

The dataset provided is the Heart Failure Prediction Dataset aggregated from the UCI Machine Learning Repository. It includes various metrics that could be predictive for heart failure from 918 patients from five different sources (Cleveland, Hungary, Switzerland, Long Beach VA and the Stalog Dataset), and a binary label for each patient, denoting if they had a heart disease or not. The data is split into 80% training set and 20% test set. The metrics provided are the following (information taken from the Kaggle page of the dataset):

- Age: Age of the patient [years]
- Sex: Sex of the patient [M: Male, F: Female]
- ChestPainType: Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: Resting blood pressure [mm Hg]
- Cholesterol: Serum cholesterol [mm/dl]
- FastingBS: Fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise]
- RestingECG: Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR: Maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: Exercise-induced angina [Y: Yes, N: No]
- Oldpeak: Oldpeak = ST [Numeric value measured in depression]
- ST_Slope: The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- HeartDisease: Output class [1: heart disease, 0: Normal]

Figure 1 shows aggregations of each of the variables of the training dataset. We note that the last plot shows that the dataset is relatively balanced. It is also worth noting that the histogram for Cholesterol seems to follow a normal distribution, except for a big spike at 0. Since a cholesterol level of 0 is close to impossible in reality we categorize these measurements as missing values for cholesterol. Upon further inspection, we see that 88% of the people in the training set with cholesterol level 0 are diagnosed with heart disease. One concerning property of the dataset is that around 80% of the patients in the dataset are males, leaving around 20% female patients. However, it was decided that it was not necessary to deal with that class imbalance.

1.1.2 Preprocessing

Due to the skewed values for cholesterol, that variable was mean imputed. After the imputation, all the numerical parameters of the dataset were scaled to have mean 0 and unit variance on the training set - the same transformation was also applied to the test set. The categorical variables were one-hot encoded for easier use in modeling. Moreover, we dropped the first column in the one-hot encoding to avoid multi-colinearity. Figure 2 shows the distributions of the preprocessed data columns belonging to the numerical parameters.

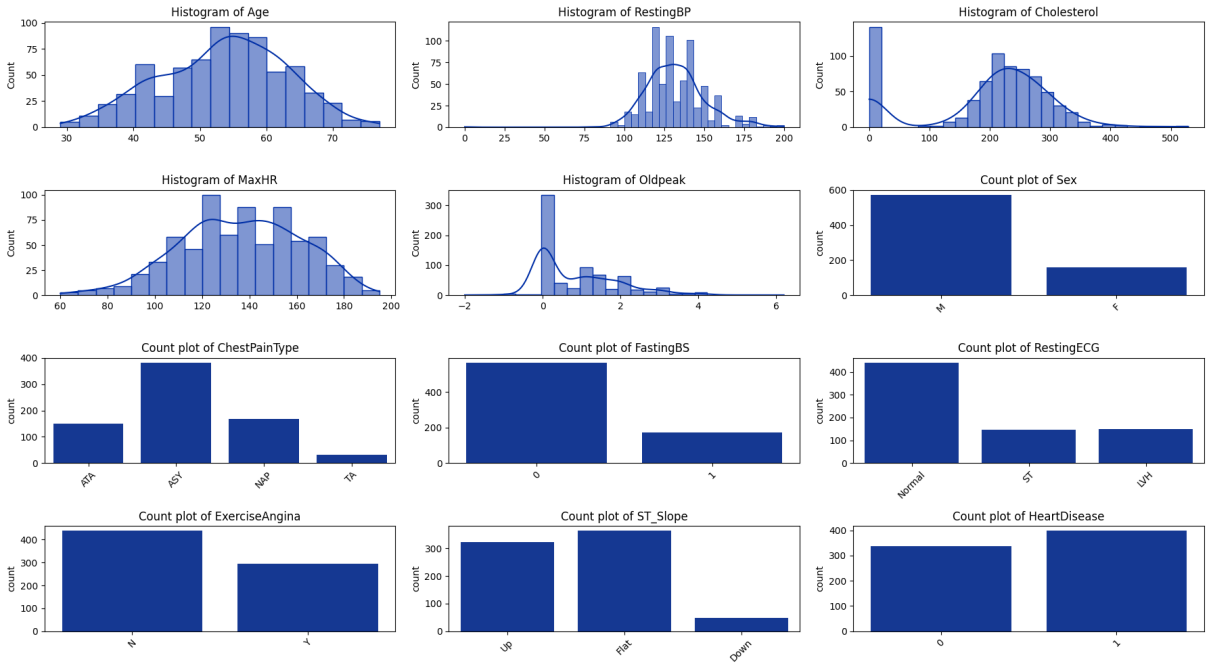

Figure 1: Aggregates and summary of each of the metrics in the dataset

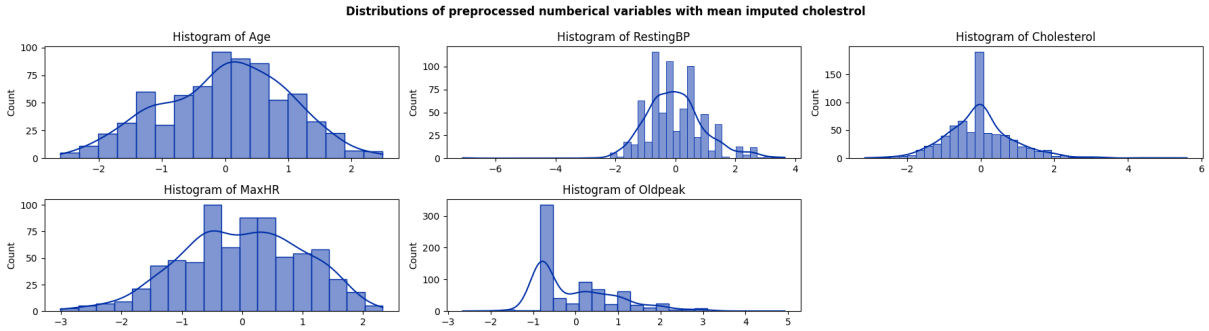

Figure 2: Aggregates and summary of the preprocessed numerical metrics

1.2 Question 2: Logistic Lasso Regression

A logistic regression model with L1 regularization, i.e. Lasso logistic regression model, was fitted to the data set. A Lasso logistic regression is a regression model trained on the dataset $\{(x_{i,1}, \ldots, x_{i,M}, y_i)\}_{i=1}^{N}$ using the following loss function:

$$\sum_{i=1}^{N} \left(-y_i \log(p_i) - (1 - y_i) \log(1 - p_i)\right) + \lambda \sum_{j=1}^{M} |\beta_j|,$$

where N is the number of observations, M is the number of explanatory variables, β_j are the model parameters, $p_i := \sigma(\sum_{j=1}^{M} \beta_j x_{i,j})$ are the predicted probabilities and λ is a hyperparameter controlling the strength of regularization.

A crucial pre-processing step to ensure that the coefficients of the model are comparable is to scale all the variables to have zero mean and unit variance. The model parameters should then all be of roughly the same magnitude, and we can say that if one model coefficient is larger than another, it indicates that this variable has a stronger relationship with the predicted quantity than the other.

The performance of the model is depicted in Table 1. Both the original dataset and the dataset with cholesterol mean imputed were compared for this regression task.

Figure 3 visualizes the model parameters. It is notable that Sex has big effect on the prediction of the model, and with further exploration, it was found that in the training set, 63% of men had heart disease, while only 24% of the women in the training set had heart disease. On the other end, we see that the three parameters corresponding to chest pain (TA, ATA, and NAP) have a big negative impact

Table 1: Lasso Logistic Regression performance on test-dataset with mean imputed cholesterol

Metric	Value
Raw Accuracy	82.0%
Balanced Accuracy	80.7%
F1 Score	85.3%

on the predicted quantity, meaning that if the patient has the fourth type of chest pain (ASY) that was dropped, the model thinks that patient is more likely to have heart disease than one with the other types of chest pain. Another notable thing to see when comparing the two images is the value of the parameter for cholesterol. When cholesterol is not mean imputed, the model has a negative coefficient in front of the cholesterol level, which should translate to people with higher values of cholesterol being less likely to get heart disease. That seems peculiar, as generally, having high values of cholesterol is labeled as bad. We see however, in the model for the data with cholesterol mean imputed, we get a positive coefficient for cholesterol level, which further supports the reasoning for mean imputing the cholesterol level in the data. In the later parts of this report, we only consider the dataset with the mean imputed cholesterol.

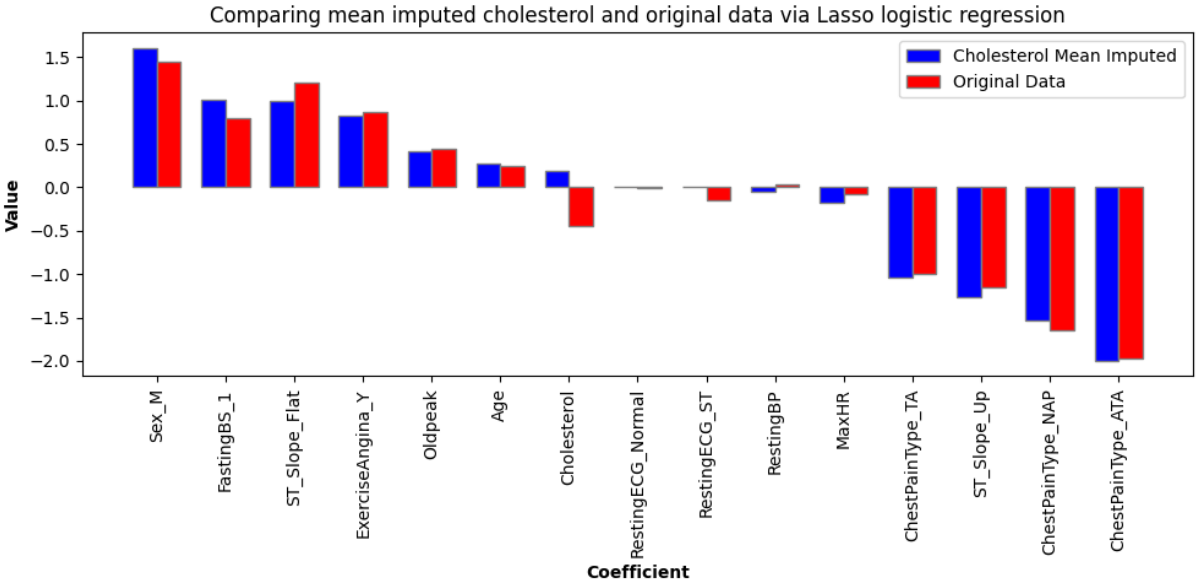

Figure 3: Visualization of model parameters without and without mean imputation for cholesterol.

Let us 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. This sounds like a good idea and is a very standard way of modeling, as it is easier to explain a model that has fewer parameters. Using Lasso logistic regression for feature selection is a widely used way to get rid of useless features, as Lasso regularization drives the parameters for useless features to exactly 0, while f.ex. L2 regularization does not. This has to do with the geometry of the unit balls of the $\|\cdot\|_1$ and $\|\cdot\|_2$ norms. For the above model we could f.ex. drop the RestingECG variable as 3 shows that for the cholesterol mean imputed data the corresponding model parameters are ≈ 0.

1.3 Question 3: Multi-Layer Perceptrons

A Multi-Layer perceptron was trained on the dataset with cholesterol mean imputed using two hidden layers, the first having 64 neurons, and the second having 32 (see Figure 4 to see a visualization of the model architecture). The neurons in the hidden layers use ReLU activation, and the neuron in the output layer uses sigmoid activation.

The performance of the model is aggregated below in Table 2.

To explain the model's prediction, Shapley values were calculated for two sample negative examples and two sample positive examples. Figures 5 and 6 contain waterfall plots for these examples, respectively.

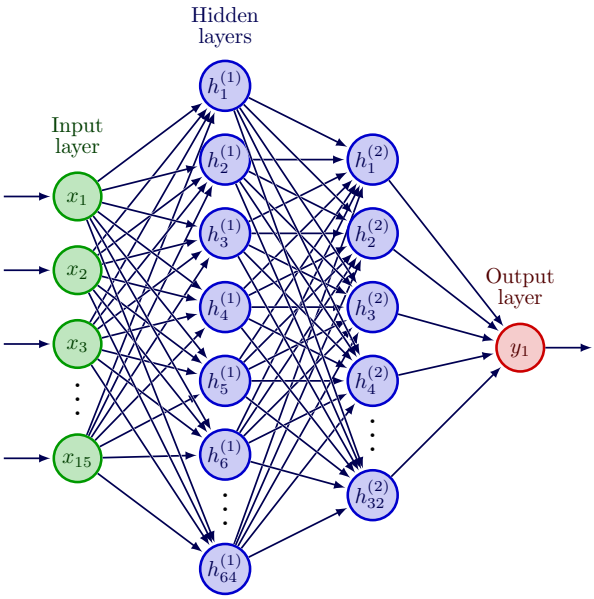

Figure 4: Visualization of the architecture of the multi-layer perceptron

Table 2: Multi-Layer Perceptron Test-Set Performance

Metric	Value
Raw Accuracy	84.7%
Balanced Accuracy	83.0%
F1 Score	87.8%

We note that the first negative example that was chosen outputs a value of 0.731, meaning it predicts that the patient has heart disease where the patient should be healthy, i.e. a false positive example. The effect of each parameter seems to be similar in every example, with small differing values, f.ex. the right plot on Figure 5 shows that since the patient has Sex_M = 1, it adds a value of 0.1 to the prediction while the patient on the left plot of Figure 6 has Sex_M = 1 and it only adds a value of 0.05 to the prediction. Besides these small numerical differences, the feature importances seem to be in line with the results from the lasso regression and there seems to be good consistency between the values, no drastic changes.

1.4 Question 4: Neural Additive Models

The Neural Additive Model from the paper provided was implemented and trained on the dataset with cholesterol mean imputed. One small deviation from the paper was to use standard ReLU activation units instead of ExU units. The reason for using ExU units according to the paper is to be able to model jagged sequences, but we argue that the effect of each parameter should be smooth. To monitor the model performance, 10% of the training data were used as a validation set, and the model that performed the best on the validation set was chosen. The performance of the model is depicted in Table 3.

Table 3: Neural Additive Model Performance

Metric	Value
Raw Accuracy	83.1%
Balanced Accuracy	82.4%
F1 Score	85.9%

Figure 7 visualizes the non-linear function fit for each model parameter. For the numerical columns, values at regular intervals between the minimum and maximum values for each of the columns were input into the model to visualize the non-linear trend of each parameter. The categorical parameters are visualized by summing the values of the one-hot-encoding for each categorical value and combined

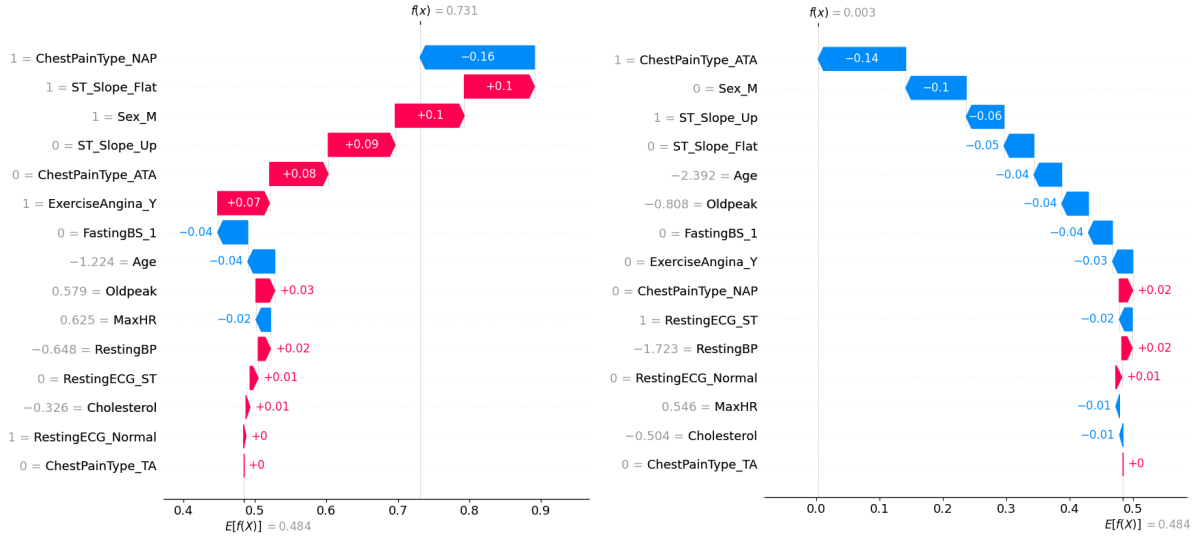

Figure 5: Waterfall plots of the Shapley values for two negative examples from the dataset, i.e. from patients that did not have heart disease

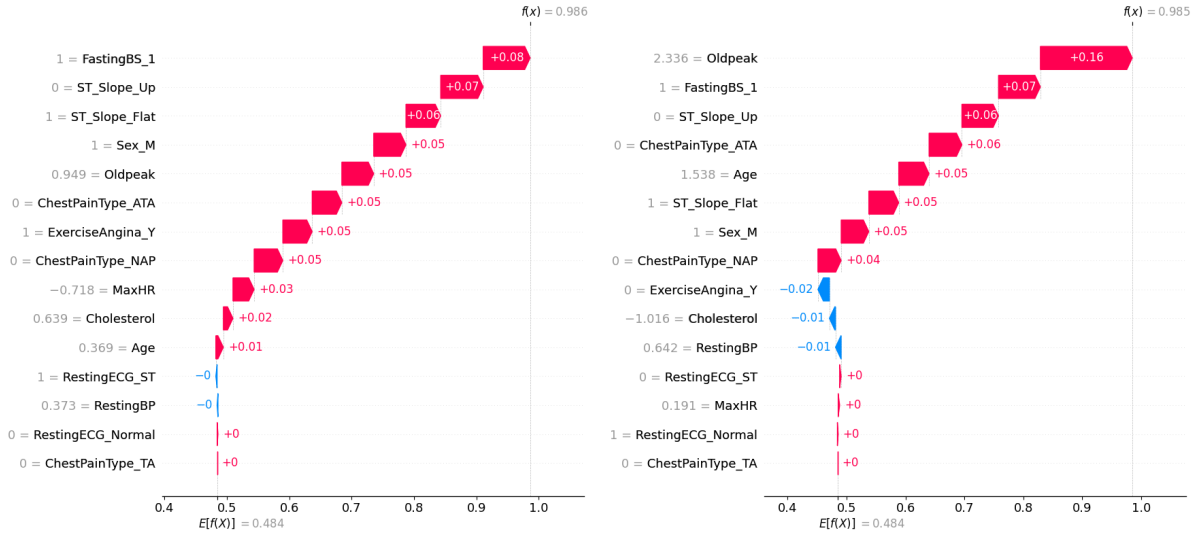

Figure 6: Waterfall plots of the Shapley values for two positive examples from the dataset, i.e. from patients that had heart disease

in the bar plots, after transforming each variable to match their original value, whereas the data that is used to train the model has each parameter scaled such that they have mean zero and unit variance. Here, a higher value means that the patient is more likely to have heart disease One interesting thing to note is that the effect of cholesterol is quite drastic. It is relatively flat until the value of cholesterol is around 200, and then increases from there. It is well known that a cholesterol level of 200 and below is considered healthy and anything above 200 is considered unhealthy.

In terms of model performance, the MLP performs best out of the three models in both metrics (BAC and F1), the neural additive model is the second best in both metrics, and the logistic regression model performs the worst. The difference is however very small in all the performance values. The small size of the dataset could be one of the factors that the neural network models are not drastically outperforming the logistic regression model.

The main benefit of Neural Additive Models over Multi-Layer Perceptrons is that we can explain how much effect each parameter has on the model output. Multi-Layer Perceptrons can have a non-linear way of combining the input variables, which makes the model hard to explain and hard to say why the model predicts what it predicts - Neural Additive Models (NAMs) are only capable of combining the non-linear activations of each parameter in a linear fashion. A NAM should also be more expressive than a logistic regression model, as it allows for non-linear trends in each of the parameters.

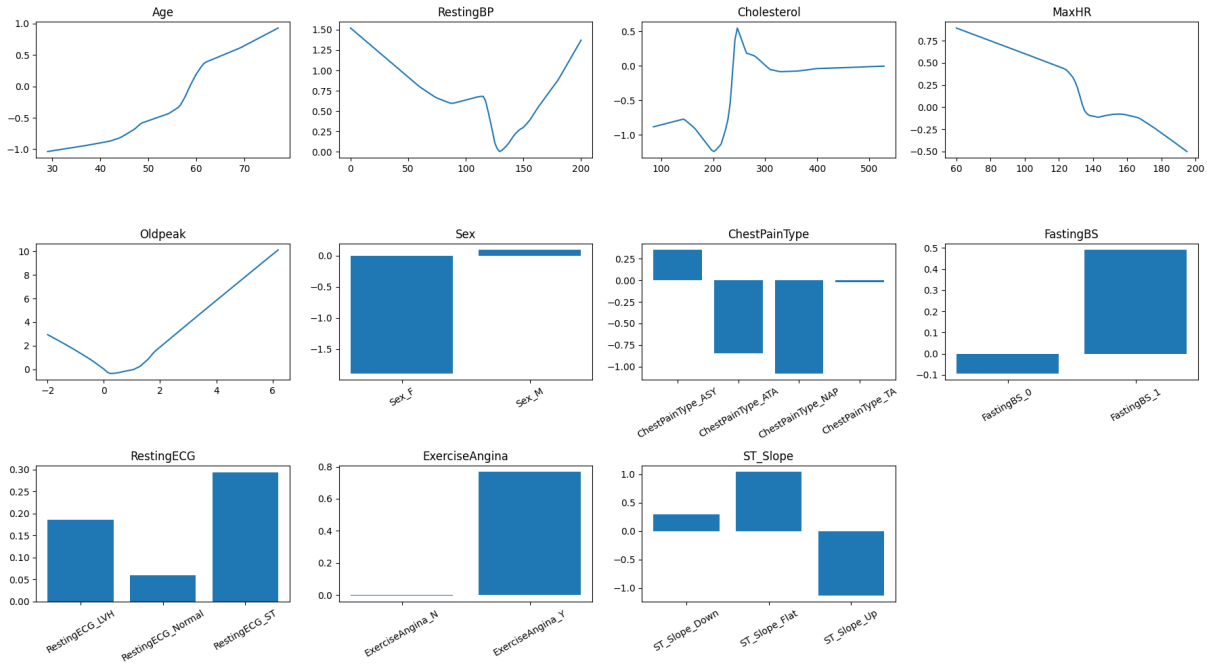

Figure 7: Visualization of the non-linear function for each of the parameters of the Neural Additive Model.

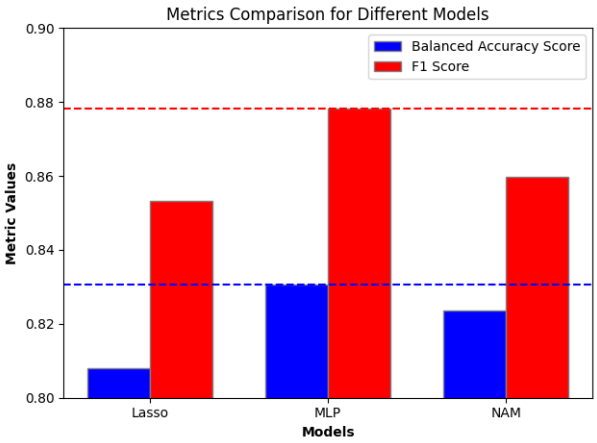

Figure 8: Model performance comparison on test-set data

2 Part 2

2.1 Question 1: Exploratory Data Analysis

The dataset provided has chest X-ray images of patients with and without pneumonia. The data is split into training data, validation data, and test data. The original data split only had 16 images for validation, so the size of that was increased to 210, 105 images in each class. Initial exploration shows that the dataset is not exactly balanced, the training set containing 3681 patients with pneumonia and only 1047 that do not have pneumonia. The test set is not balanced either, including 234 normal patients and 390 patients with pneumonia. Figure 9 shows the label distribution for the training, validation, and test sets. Figure 10 shows examples of images of healthy patients and patients with pneumonia.

We see some notable differences in the two types of images. The lungs of the patients with pneumonia appear more cloudy and fuzzy, in contrast to the lungs of the healthy patients being of sharper/clearer quality. However, for some of the images, it is not visually clear which class they belong to.

One potential source of bias that could influence model performance is the data is not balanced. That can be mitigated by looking at other metrics than just accuracy, such as balanced accuracy or F1-score. In addition to that, for some patients the dataset contains more images than for other patients - this

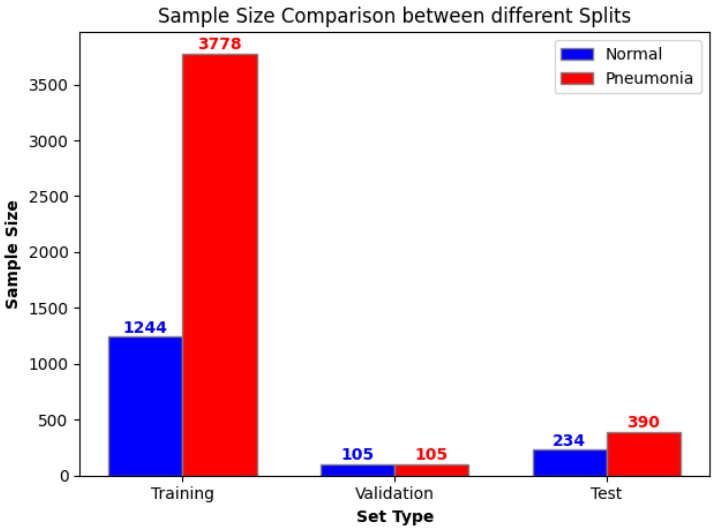

Figure 9: Label distribution of the training, validation and test sets.

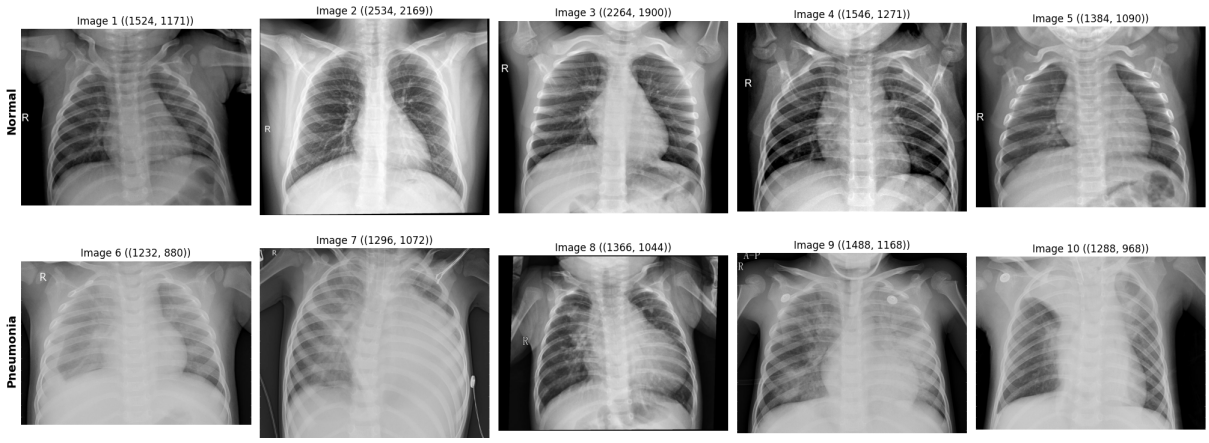

Figure 10: Examples of raw X-ray images from the training dataset. The top row shows examples of normal patients and the bottom row shows examples of patients with pneumonia

imbalance could lead to an additional model bias.

2.1.1 Data Pre-Processing

The images were all rescaled to a size of 240×240 pixels in size. Moreover, we normalized the RGB intensities by $\frac{1}{255}$, s.t. they yield values in the range from 0 to 1 - this helps to accelerate convergence in the training phase and makes computations more stable. In addition to that, we augmented the training data by random shearing and zooming.

2.2 Question 2: CNN Classifier

A CNN model with the architecture shown in Figure 11 was fitted to the training data. The architecture consists of three successive convolutional blocks, each consisting of a convolutional layer (3x3 kernel) followed by a max-pooling layer (2x2 kernel). The convolutional blocks are then followed by three blocks each consisting of a dense layer followed by a dropout layer. The final output is computed by passing the activation through a sigmoid function, yielding a result between 0 and 1.

The model performance can be found below in Table 4

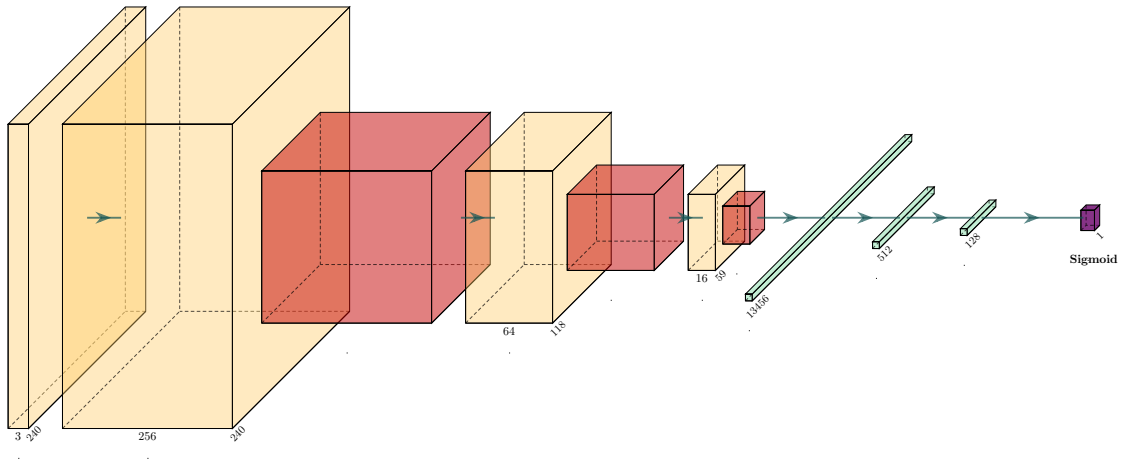

Figure 11: Graphical representation of the CNN architecture. The yellow boxes depict convolutional layer, each having a 3×3 kernel, padding to match the size of the image, and ReLU activation. The red boxes depict a max pooling layer followed by a batch normalization layer. All the max pooling layers have 2×2 kernels and padding to match the size of the image. The dense layers in the end have dropout with probability $p = 0.5$ to prevent overfitting.

Table 4: Test set performance of the CNN pneumonia classifier

Metric	Value
Raw Accuracy	89.9%
Balanced Accuracy	86.9%
F1 Score	91.4%

2.3 Question 3: Integrated Gradients

The integrated gradients post-hoc method was implemented and is visualized in the images on Figure 12. In each of the plots, we can see the baseline image that was used (in this case a black image), the original image, the attribution mask of the algorithm and the attribution mask overlayed on the original image to visualize the parts of the images that are highlighted.

We see that the intensity of the pixels on the image seem to vary quite drastically from image to image, and there is not a very consistent pattern in the attribution masks. We see in some images, that the attribution mask seems to highlight the rib cage and slightly highlight the lungs behind the patients rib cage, which is sensible. For some reason, the attribution masks highlight the outline of the rib cage, which does not seem like the right thing to look at.

We see that the attribution masks are consistent across samples in the sense that they highlight similar regions for both healthy patients and pneumonia patients, but the highlighting colors differ slightly. For the examples of the patients that have pneumonia, the integrated gradients algorithm highlight the lung in a brighter color and the outline of the ribs in a darker color, while the opposite goes for the healthy patients. This is because the gradient steps are in different directions, i.e. when the true label is Normal the gradients should go in one direction, and when the true label is Pneumonia the gradients should go in another direction.

Figure 13 compares two different baseline images, a black image and a uniformly sampled grayscale image. We see that the attribution masks corresponding to the black baseline seem to highlight sensible regions while the ones corresponding to the uniform baseline seem to be way more random, highlighting a similar area but in a more fuzzy manner. Thus, in the later parts of this report, we will only look at black baseline images.

2.4 Question 4: Grad-CAM

The Grad-CAM post-hoc method was applied to the trained CNN model and Figure 14 shows a visualization of the areas of focus of the last convolutional layer of the model for 5 healthy patients and 5

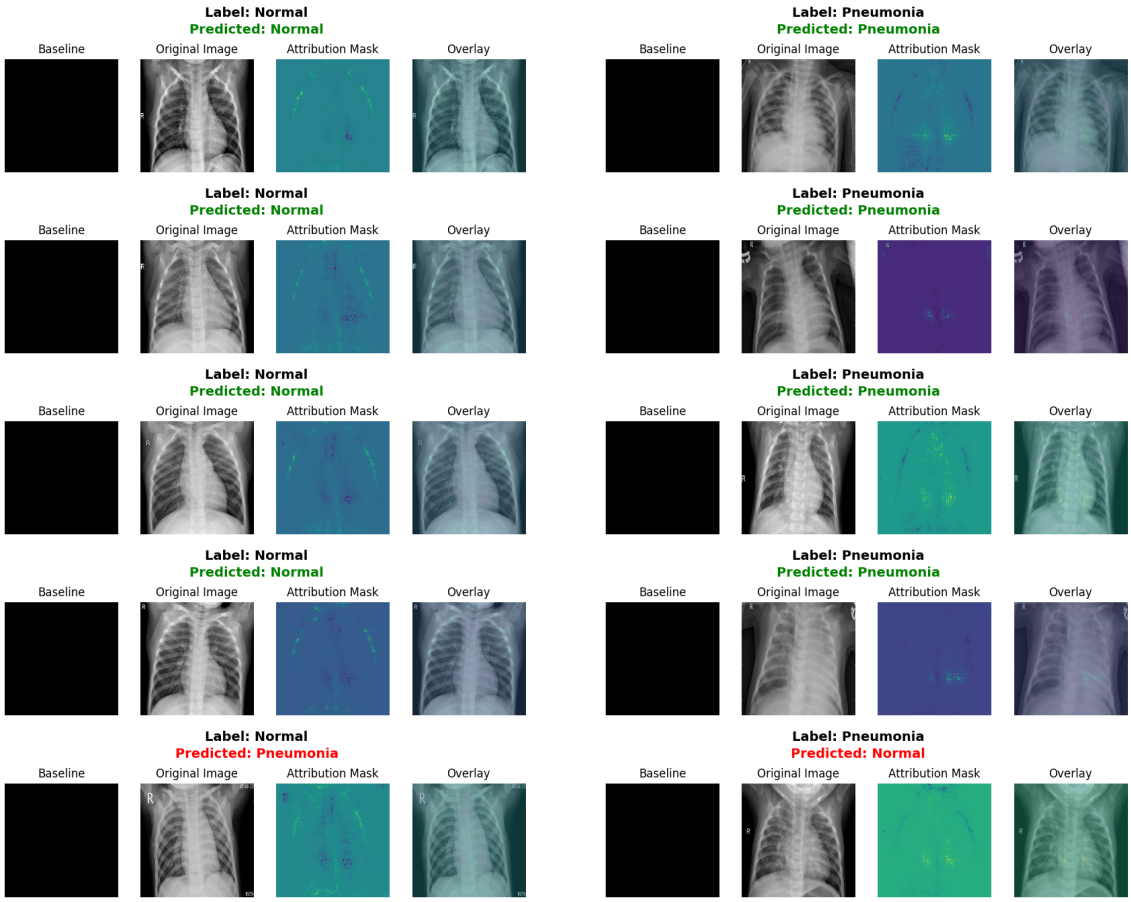

Figure 12: Visualization of the integrated gradients with black baseline image - the overlay was computed using $\alpha = 0.4$

patients with pneumonia.

We note that the Grad-CAM method highlight the lung, which is sensible, but it also highlights the outline of the rib cage and the spine behind the lungs quite intensively, which seems slightly incorrect. Since we are no experts, we note that this might also be an important region of the lung to search for pneumonia.

The attributions of the Grad-CAM seem to be quite consistent between healthy patients and pneumonia patients, highlighting similar areas (as described above).

Comparing the attribution masks for the integrated gradients and the Grad-CAM methods, we see that both methods highlight both the lung and the outline of the rib cage. The integrated gradients method however only seems to highlight part of the lung while Grad-CAM highlights the whole lung, which feels a bit more trustworthy.

2.5 Question 5: Data Randomization Test

A CNN with the same architecture as depicted in figure 11 was trained on the dataset, but with the labels of the examples shuffled. The performance of the model is depicted in Table 5. We note that the balanced accuracy of the model is exactly 50%, and within further inspection, we saw that the model predicts that every person in the test set has Pneumonia, meaning that the model was not even able to overfit to the training data. This could be for example be because of overfitting prevention techniques used in the architecture such as dropout.

The integrated gradients method and Grad-CAM were applied in the same manner as in the previous parts, and Figures 15 and 16 show the results for those two methods respectively. We see that for the integrated gradients, the model seems to very randomly highlight areas of the image, especially when looking at the overlayed images, meaning that the model did not learn any useful features. For the Grad-CAM images, the model seems to almost highlight the areas of the images proportionally to their pixel intensity, which also supports the claim that the model did not learn any useful features.

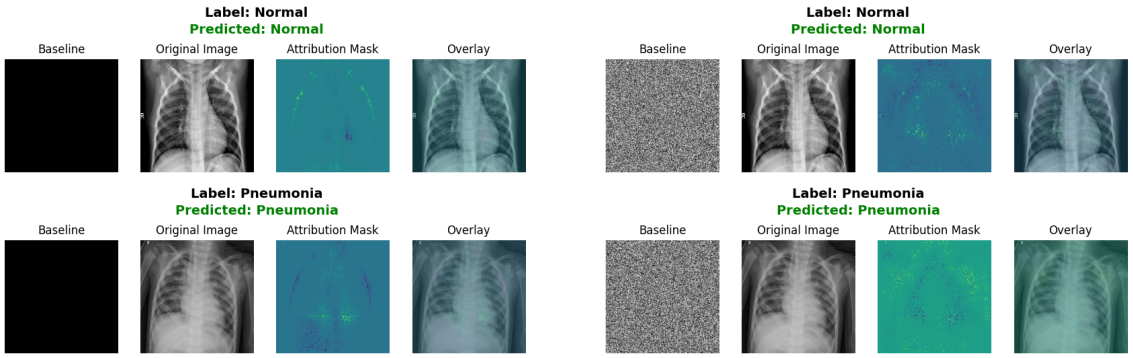

Figure 13: Comparison of integrated gradients attribution maps for black and uniformly random baseline image - the overlay was computed using $\alpha = 0.4$

Table 5: Test set performance of the CNN pneumonia classifier with the labels randomized

Metric	Value
Raw Accuracy	62.5%
Balanced Accuracy	50.0%
F1 Score	76.9%

Judging from the results explained above, we conclude that both of the methods pass the data randomization test, i.e. not being able to capture any useful features in the data when the labels are permuted randomly.

3 Part 3

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.

3.1 Question 1

3.1.1 Part 1

Comparing the three methods for the heart disease prediction, Logistic Lasso Regression, Shapley values, and Neural Additive Models, we see that all of them have similar trends for each of the features. For example, the ASY chest pain type seems to be a highly determining factor of predicting heart disease for all of the methods. The NAM produces slightly different patterns to those produced by the logistic regression model in the sense that it can have a non-linear trend for each parameter, which is clearly seen from the feature importance of cholesterol.

3.1.2 Part 2

Comparing the attribution masks for the integrated gradients and the Grad-CAM methods, we see that both methods highlight both the lung and the outline of the rib cage. The integrated gradients method however only seems to highlight part of the lung while Grad-CAM highlights the whole lung.

3.2 Question 2

3.2.1 Part 1

The NAM would be relatively easy to explain to a doctor, since it is possible to visualize the trends for each variable in a very explanatory manner, and how much each factor attributes to the prediction. We could argue that this type of method would speed up their analysis of each patient drastically, freeing up time for other patients or other things. The NAM model can also output the confidence score for each example, which could help determining which examples need a closer look than others. The feature trends depicted in Figure 7 could help arguing that the model understands which features are important for determining if a patient has heart disease or not.

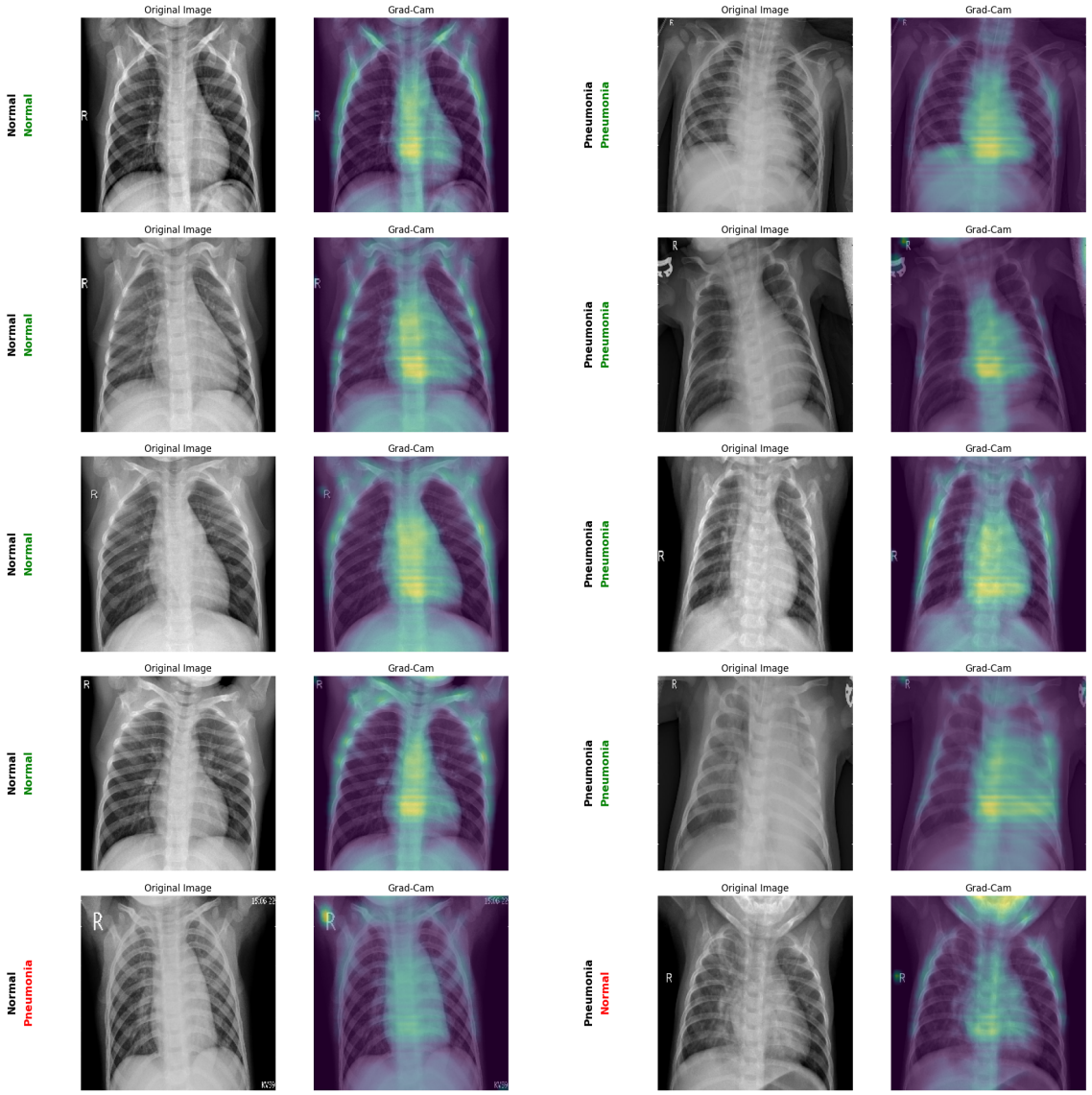

Figure 14: Visualization of the Grad-CAM method for 5 healthy patients and 5 patients with pneumonia

3.2.2 Part 2

Seeing as CNNs are generally black-box models and no one can really explain why they predict what they predict, it could be hard to convince a doctor to use these models. We could however point to the fact that the accuracy is quite high (almost 90%) and that it could reduce the time it takes for them to determine if a patient has pneumonia or not. We could also argue that the Grad-CAM method highlights somewhat sensible regions and show them confidence scores (prediction probabilities) for each prediction to aid in decision making.

3.3 Question 3

3.3.1 Part 1

For heart disease prediction, the explainability methods all seemed to place a big emphasis on the type of chest pain that the patients have, which is very natural for predicting heart disease, as pain in the chest area is intuitively a symptom of heart disease. Gender was also an important feature for each of the explainability methods, which seems non-sensible but is likely caused from the fact that 24% of women in the data had heart disease while 63% of the men did, so the models seem to learn that women are less likely to have heart disease, which in our opinion should generally not necessarily be true.

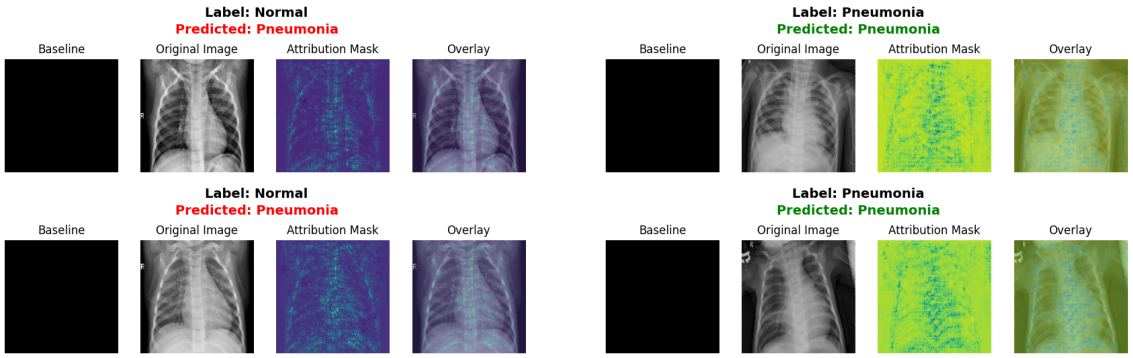

Figure 15: Visualization of the integrated gradients with black baseline image for data with permuted labels - the overlay was computed using $\alpha = 0.4$

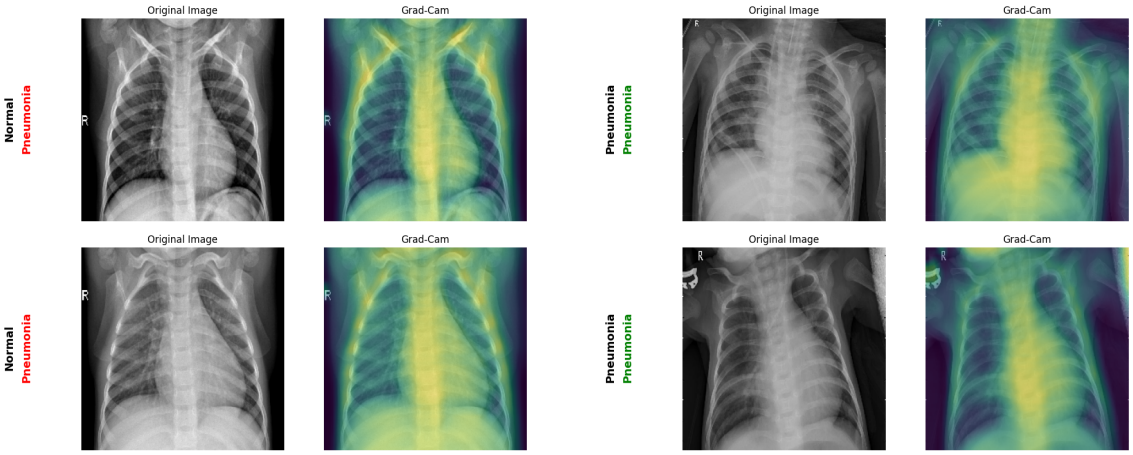

Figure 16: Visualization of the Grad-CAM method for the model trained on permuted labels for 5 healthy patients and 5 patients with pneumonia

3.3.2 Part 2

Both the integrated gradients and the Grad-CAM methods highlighted the lungs, which is sensible for predicting lung disease like pneumonia. However, both of them also highlighted the outline of the rib cage as important features, which seems totally non-sensible. Maybe it would help in determining how the lung is oriented compared to the rib cage, but that seems like a far-fetched explanation of this as the rib cage should not be very important in predicting a lung disease like pneumonia.

3.4 Question 4

3.4.1 Part 1

Among the three models considered in part 1, the NAM model demonstrates the best balance between interpretability and performance. It outperforms Logistic Regression by allowing for non-linear trends for each parameter while preserving interpretability. Compared to MLP, it performs slightly worse due to its inability to combine parameters in a non-linear fashion. However, MLPs lack interpretability. Therefore, the NAM model is the most suitable choice for a deployment in practice.

3.4.2 Part 2

Comparing the Grad-CAM and integrated gradients methods, we can see that both show a consistent behavior in highlighting the patient's lungs - but the Grad-CAM highlights the whole organ, while the integrated gradients only highlights small spots in the lung area. However, both methods also highlight regions, such as the edge of the rib cage, that do not seem to make sense to look at when faced with the task of diagnosing pneumonia. If we had to choose one method over the other, it would be the Grad-CAM because the highlighting is simply more expressive - however, neither method leads to a completely satisfactory result.

4 Project Code

All the code including images produced for the project can be found in the GitHub repository for the project, but the notebooks, README and requirements.txt were handed in as .zip files to Moodle