Project 1 Report: Interpretable and Explainable Classification for Medical Data

Mateo Boglioni, Federica Bruni, Paula Momo Cabrera

April 16, 2024

1 Part 1: Heart Disease Prediction Dataset

This project aims at developing interpretable and explainable classification models for medical data, particularly focusing on predicting heart disease outcomes.

1.1 Exploratory Data Analysis

Firstly, we conducted a comprehensive analysis of the features in our training dataset and their distributions. Our dataset consists of 734 samples representing 12 categories, both numerical and categorical. An initial investigation of the dataset immediately reveals that it contains no null values and is therefore complete. We then conducted a detailed analysis, first focusing on the numerical features and then on the categorical ones.

Numerical Features

There are 7 numerical features, which are listed in the table below together with the interval of their values.

Feature	Description	Interval
AGE	age of the subjects	[29, 77]
RESTING BP	resting blood pressure of the patient	[0, 200]
CHOLESTEROL	cholesterol level of the patient	[0, 530]
MAX HR	maximum heart rate achieved by the patient	[60, 195]
OLD PEAK	ST depression induced by exercise relative to rest	[-2, 6.2]

Table 1: Numerical Features

Then we proceeded to analyze the distribution of the features (Figure1). Age, resting blood pressure (RestingBP), and maximum heart rate (MaxHR) exhibit a normal Gaussian distribution. Similarly, the cholesterol variable demonstrates an underlying normal distribution, albeit with 19% of samples displaying a cholesterol level value of 0. We suspect this may be due to misreporting and consider these instances as potential missing values rather than true zero values.

These insights guided our decision regarding data normalization or standardization before further processing. We adopted both normalization and standardization.

- **Normalization:**

  Normalization was applied to features with non-normal distributions, such as the Oldpeak feature, which exhibited a right-skewed distribution and was normalized using MinMaxScaler.

- **Standardization:**

  Standardization was performed on features with normally distributed data but with values that significantly deviated in scale from other features. Specifically, age, restingBP, and MaxHR features were standardized using StandardScaler, while the cholesterol feature was scaled using RobustScaler due to its sensitivity to outliers.

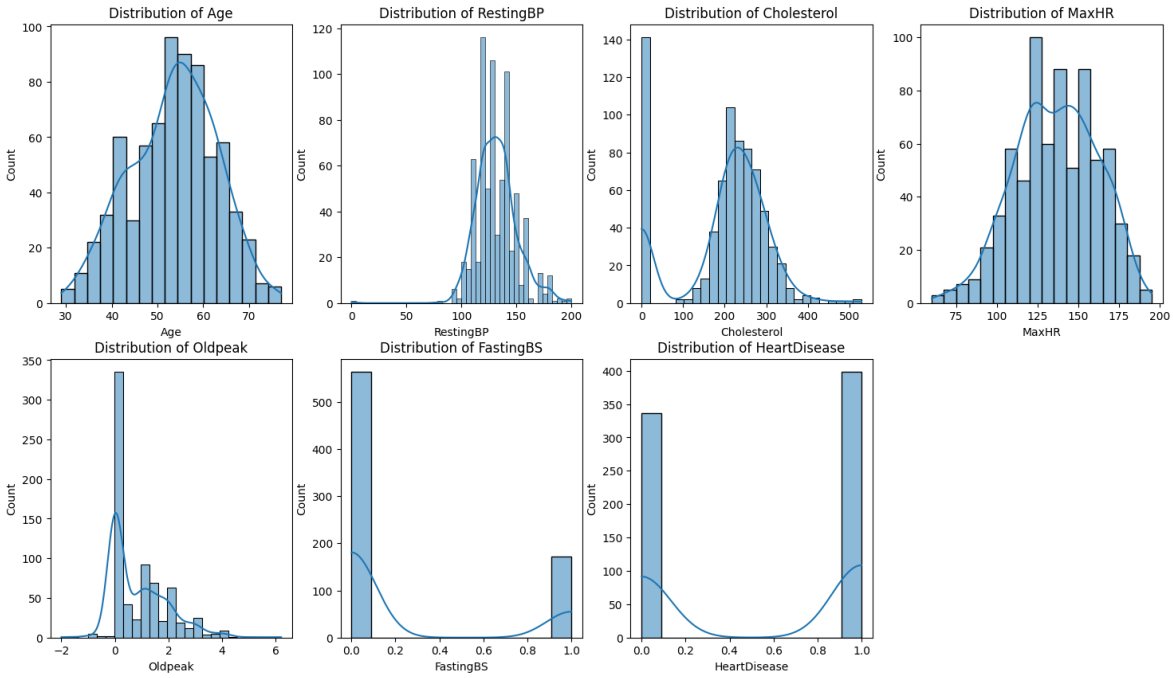


Figure 1: Histograms displaying count distribution of numerical features

Categorical Features

In our dataset, we have 5 categorical features, each accompanied by its respective encoding, as presented in the table below Furthermore, to facilitate further processing and modeling, categorical values

Feature	**Description**	**Encoding**
SEX	gender of the patient	{'Female (F)': 0, 'Male (M)': 1}
CHEST PAIN TYPE	type of chest pain experienced by the patient	{'Asymptomatic (ASY)': 0, 'Atypical Angina (ATA)': 1, 'Non-anginal pain (NAP)': 2, 'Typical Angina (TA)': 3}
RESTING ECG	electrical activity of the heart while at rest	{'Left ventricle hypertrophy (LVH)': 0, 'Normal': 1, 'ST segment depression (ST)': 2}
EXERCISE ANGINA	chest pain occurrence during physical activity	{'No (N)': 0, 'Yes (Y)': 1}
ST SLOPE	direction and angle of the ST segment	{'Down': 0, 'Flat': 1, 'Up': 2}

Table 2: Categorical Features

Sex, type of chest pain (ChestPainType), Resting 12-lead electrocardiography(RestingECG), exercise-induced chest pain (ExerciseAngina) and ST_Slope [KBM24](referring to the ST segment portion of the electrocardiogram (ECG) waveform that represents the interval between ventricular depolarization and repolarization), were encoded into numerical representations as shown in Table 2.

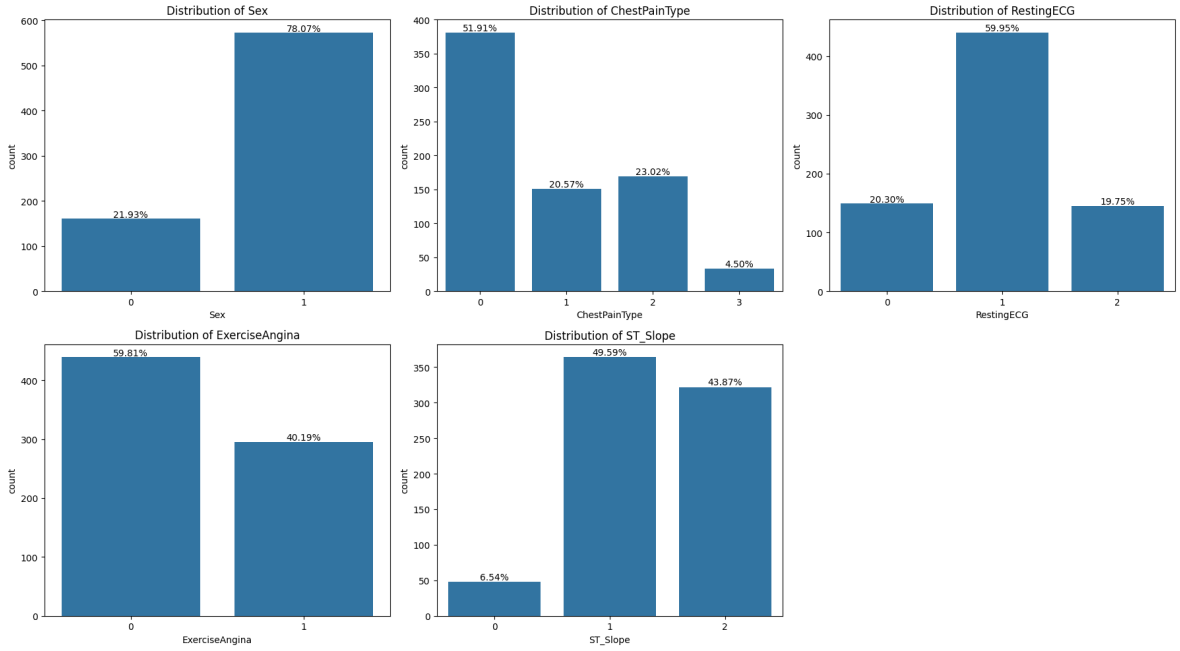


Figure 2: Categorical features distribution

Then we analyzed the dataset's balance. The pie chart below (figure 3) shows that 45.8% of individuals have heart disease (1), while 52.4% do not report a heart condition (0), indicating a balanced dataset.

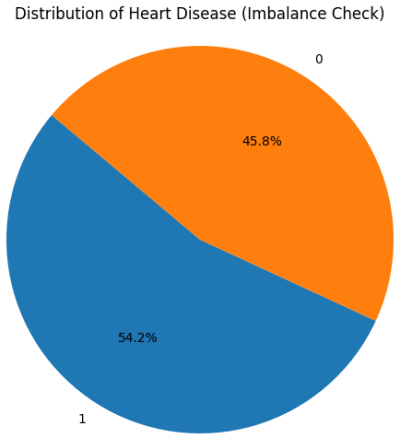


Figure 3: Overall dataset distribution of Heart Disease

Prevalence of heart disease is known to vary across different demographic and physiological factors [YR01]. Notably, in our dataset, individuals aged over 55, with resting blood pressure exceeding 130, and a maximum heart rate capacity below 130, displayed higher rates of heart disease. Those exhibiting an Oldpeak value greater than 1 and testing positive for fasting blood sugar also showed increased prevalence. Furthermore, a significant disparity was observed between males and females, with over 49% of males compared to 5% of females presenting heart disease. Sex-based differences in heart disease manifestation and diagnosis have been consistently reported, with males historically exhibiting higher prevalence rates [FBM+23]. Therefore, the unequal representation of sexes in our dataset may impact model accuracy. Additionally, individuals experiencing exercise-induced angina had a higher prevalence of heart disease compared to those who did not 34% *vs.* 20%). Surprisingly, individuals with asymptomatic chest pain also displayed a higher prevalence of heart disease. Lastly,

individuals reporting a flat ST_slope showed a heart disease prevalence of 34%.

These observations were made based on the following graphs, which depict the distribution of features in comparison with the target class.

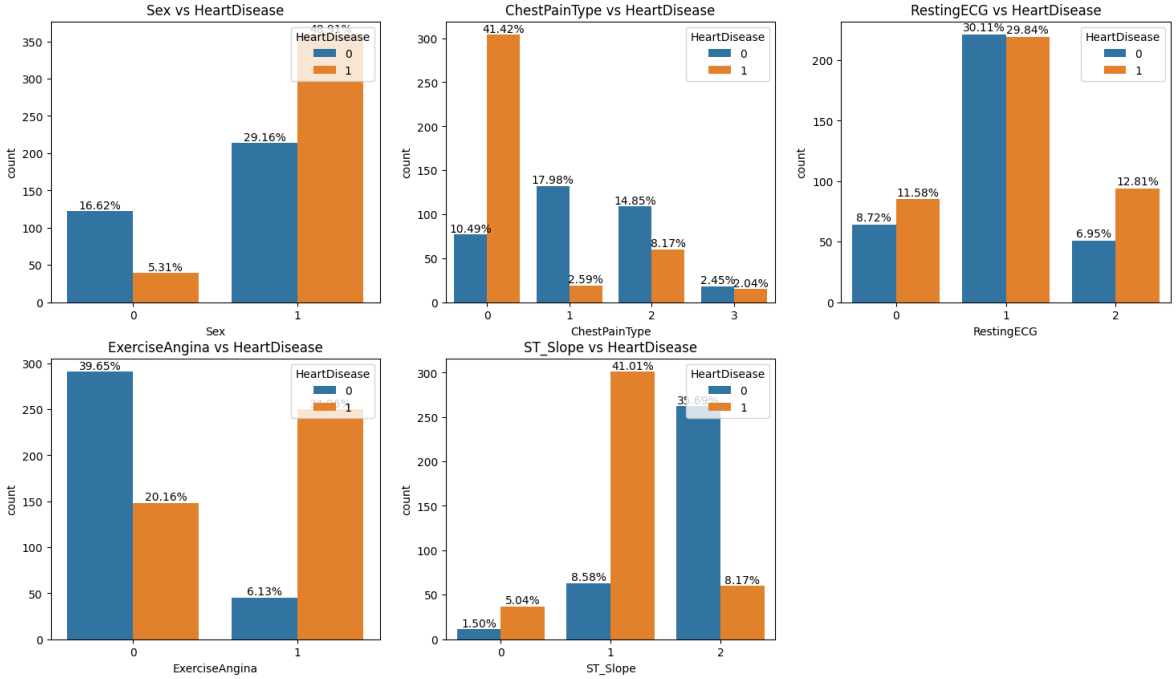


Figure 4: Categorical Features Distribution vs Heart Disease

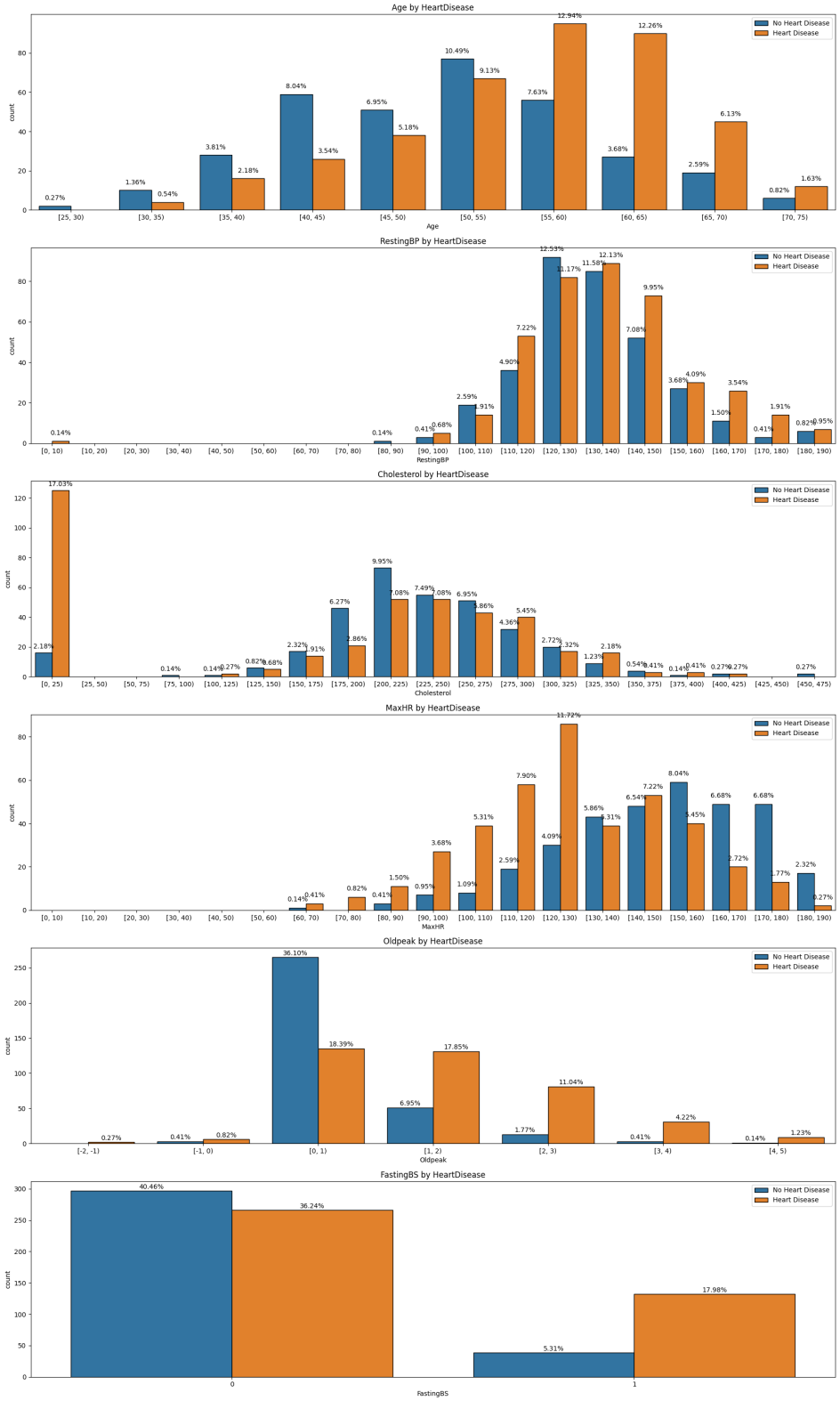


Figure 5: Numerical Features Distribution vs Heart Disease

1.2 Logistic Lasso Regression

In fitting a Lasso regression model to the combined heart disease dataset, essential pre-processing steps ensure feature coefficient comparability. Standardization or normalization prevents features with larger magnitudes from dominating regularization. Missing value handling, either through imputation or deletion, prevents adverse effects on model performance and interpretation. Feature selection removes redundant or irrelevant features, reducing dimensionality and enhancing interpretability.

To determine the best alpha value, LassoCV with 7-fold cross-validation is used, minimizing mean squared error and improving performance. The model is then fitted to the training data with the optimal alpha. Performance evaluation involves computing scores for both training and test data, quantifying the proportion of variance explained by the model. Predicted probabilities are converted to binary predictions using a threshold of 0.5.

The Lasso regression model, with a best alpha of 0.000202, demonstrates effective regularization, yielding training and test scores of 0.532 and 0.425, respectively. The high F1-score of 0.870 indicates accurate classification of heart disease outcomes.

Further assessment involves calculating the F1-score, providing balanced accuracy insights. Visualization of feature importance aids interpretation by highlighting each feature's contribution to the model's output, facilitating decision-making.

The confusion matrix based on the F1-score results provides valuable insights into the model's performance. Upon analysis, it is observed that 97 out of 113 (86%) true positives were correctly predicted as positives, while 58 out of 71 (82%) true negatives were correctly predicted as negatives. These results indicate a relatively high level of agreement between the model's predictions and the true labels in both positive and negative classes. The high percentage of correctly predicted true positives suggests that the model effectively identifies instances with heart disease, while the high percentage of correctly predicted true negatives indicates accurate identification of instances without heart disease (Figure 6).

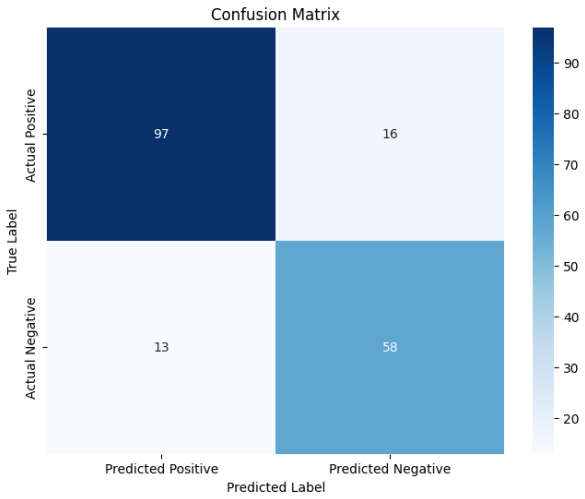


Figure 6: Confusion matrix from Lasso Regression F1

Incorporating both the pre-processing steps and the results from the model evaluation process enhances the overall assessment of the Lasso regression model's performance. By examining feature coefficients, performance metrics, and the distribution of true positives, false positives, true negatives, and false negatives, stakeholders can gain a comprehensive understanding of the model's predictive capabilities and make informed decisions regarding heart disease outcomes.

The results of the Lasso regression provide valuable insights into the relative importance of each feature in predicting heart disease outcomes. Among the features, Oldpeak emerges as the most influential, followed by ExcersiseAngina with a positive coefficient indicating that higher values are associated with an increased likelihood of heart disease. Interestingly, these results aligning with their strong positive correlations with heart disease when performing Pearson correlation analysis (Figure 4). Conversely, features like ST_Slope and ChestPainType exhibit negative coefficients, suggesting certain patterns in the ST segment of the ECG and specific types of chest pain may be indicative of a lower

likelihood of heart disease. Other factors such as ExerciseAngina, sex, and FastingBS show positive coefficients, indicating associations with increased risk. Notably, while age and cholesterol exhibit relatively small coefficients, they still contribute to the predictive power of the model. These findings underscore the multifactorial nature of heart disease risk, highlighting the importance of considering various clinical markers in assessing cardiovascular health.

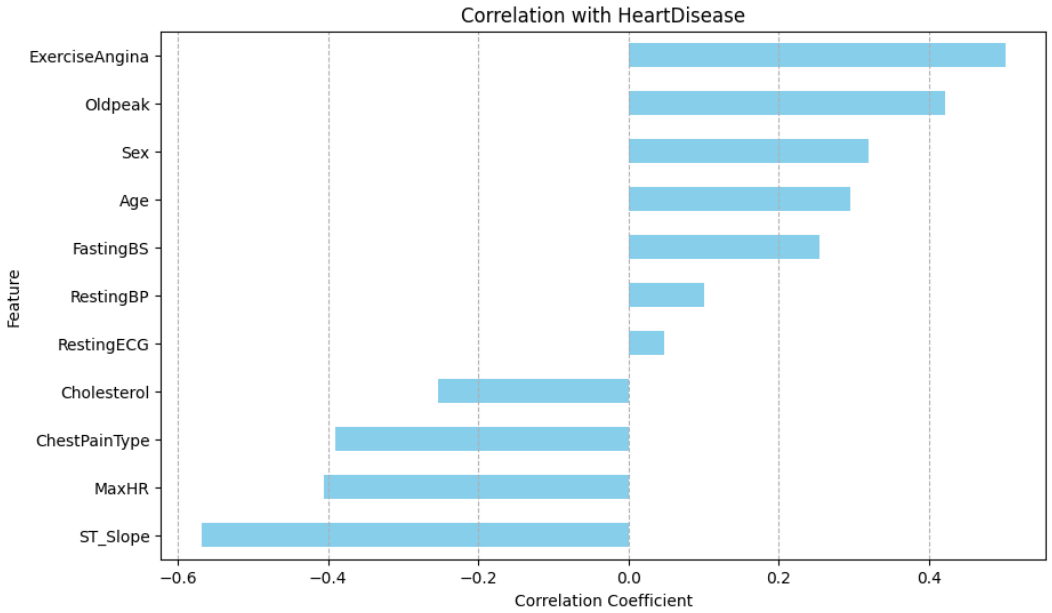


Figure 7: Feature correlation with Heart Disease

Utilizing Logistic Lasso Regression for feature selection followed by training a Logistic Regression solely on these variables presents both advantages and drawbacks. On the positive side, Logistic Lasso Regression effectively applies L1 regularization to promote sparsity in feature coefficients, aiding in feature selection and enhancing model interpretability. It simplifies the model, reduces overfitting, and may improve generalization by prioritizing informative features. However, training a Logistic Regression solely on the selected variables may overlook relevant features excluded during selection, potentially leading to loss of information and biased conclusions. Additionally, the simplified model may not capture the data's complexity adequately, and its performance may be less robust than the original model. Thus, while this approach offers benefits such as enhanced interpretability and potential performance improvements, careful consideration of its impact on predictive accuracy and reliability is essential. Researchers should weigh the advantages of simplicity against the drawbacks of feature exclusion based on the dataset's context and research objectives.

1.3 Multi-Layer Perceptrons

Multilayer Perceptrons (MLPs) are renowned for their superior performance in various machine learning tasks [noa]. However, their interpretability poses a significant challenge, hindering our ability to comprehend their decision-making process. To overcome this limitation, we employed post-hoc explainability methods such as SHAP (SHapley Additive exPlanations), which provide insights into feature importance and model behavior [SSG+24].

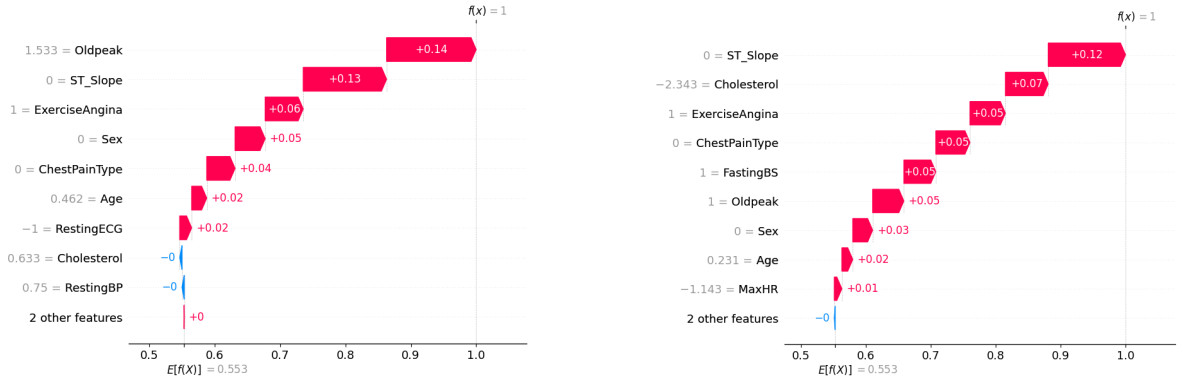
In our methodology, we initiated by training a simple MLP classifier on a dataset pertaining to heart failure. The dataset underwent preprocessing steps, including encoding categorical features and scaling numerical features using a robust scaler. Following this, we trained the MLP classifier with optimized parameters for classification tasks, aiming to predict heart disease outcomes.

The results of our approach were promising, with the trained MLP classifier achieving a test set accuracy of 82.61%. This accuracy demonstrates the model's ability to capture patterns in the data and make accurate predictions.

To enhance interpretability and understand the inner workings of the MLP classifier, we leveraged SHAP explanations. By visualizing SHAP explanations for individual predictions, we gained valuable

insights into the importance of each feature in determining model predictions. This approach allowed us to interpret the model's decisions and understand its behavior, despite the inherent complexity of MLPs. We utilized the SHAP library to compute SHAP values, assessing the influence of features on the MLP model's predictions regarding heart disease. Then, we selected two positive and two negative samples based on the model's predictions aligned with the true labels for further examination.

When examining the two randomly selected positive samples for heart disease, a consistent pattern emerged where certain features tended to elevate the likelihood of the model predicting heart disease. Notably, attributes such as ST_Slope, ExerciseAngina, Oldpeak, and ChestPainType consistently exhibited the highest positive contributions to the model's heart disease prediction in both samples. In this case, a ST_Slope of 0 is indicative of an downward slope in the ST segment of an ECG, which is often linked with hypokalemia, myocardial ischemia, and a left bundle branch block [KBM24], thus predisposing individuals to heart disease. Similarly, factors like ExerciseAngina, which denotes chest pain or discomfort during physical exertion, and Oldpeak (ST depression induced by exercise relative to rest), are recognized indicators of cardiac stress and potential coronary artery disease [LMS+04]. The presence of different types of chest pain, as captured by ChestPainType, can indeed offer valuable insights into the likelihood of heart disease, as specific patterns may strongly correlate with cardiac issues. However, the challenge arises from the numerical encoding of this variable, which encompasses multiple responses without clear gradations of severity or risk. Assigning numerical values to these responses in a predictive model may introduce confusion, as it's difficult to establish a meaningful ranking or hierarchy among them. This lack of clear gradation could potentially confound the model's predictions. Overall, the observed positive SHAP values for these features reinforce their clinical relevance in predicting heart disease and underscore the utility of the model's interpretations in aligning with established medical knowledge.

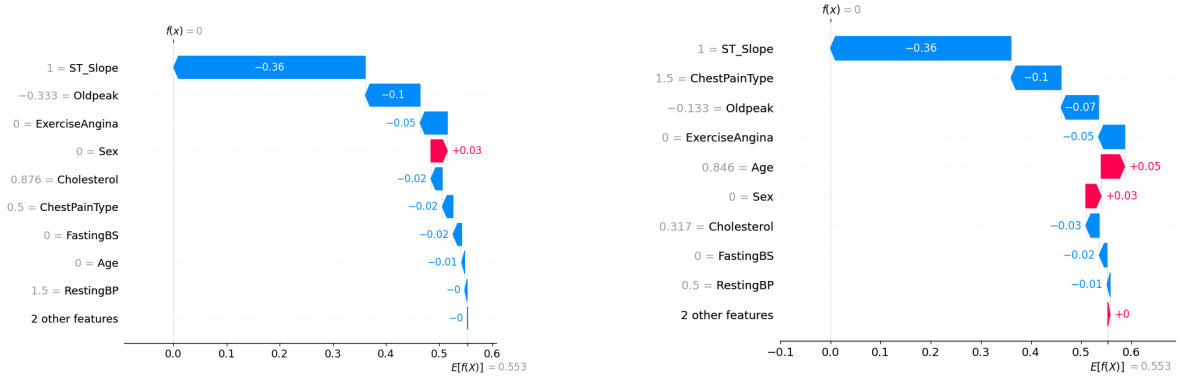


(a) Sample 1 (b) Sample 2

Figure 8: SHAP model feature contributions using samples positive for Heart Disease

Upon reviewing the two randomly selected negative samples for heart disease, a consistent trend emerged, indicating that specific features negatively impacted the model for predicting heart disease. Notably, a ST_slope of 1, which indicates a "flat" slope, consistently emerged as the most significant predictor of a lower risk of heart disease, followed by diminished values of Oldpeak and ExerciseAngina, among others. This pattern suggests that individuals exhibiting these characteristics are less likely to be diagnosed with heart disease according to the model's predictions. Clinically, this aligns with expectations, as a shallower or flatter ST segment on an electrocardiogram (ECG), is generally associated with a lower risk of myocardial ischemia or infarction [HGS+10], thus reducing the likelihood of heart disease. Similarly, lower values of Oldpeak, reflecting a smaller magnitude of ST depression during exercise relative to rest, and reduced incidence of ExerciseAngina, are consistent with lower cardiac stress and decreased probability of coronary artery disease [LMS+04]. The observed pattern underscores the clinical relevance of these features in assessing heart disease risk and emphasizes the interpretability of the model's predictions in the context of established medical knowledge.

The consistent patterns observed in both positive and negative samples for heart disease across the entire test dataset confirm the robustness of the model's predictions. Higher values of features like FastingBS, Oldpeak, cholesterol, age and MaxHR consistently increased the likelihood of heart disease



(a) Sample 1 (b) Sample 2

Figure 9: SHAP model feature contributions using samples negative for Heart Disease

prediction, aligning with clinical expectations. Conversely, lower values of these features were associated with a reduced likelihood of heart disease prediction, reflecting established medical knowledge regarding cardiac health indicators. Notably, high values of ST_slope associates with less likelihood of heart disease (Figure 10). These findings highlight the model's ability to interpretably align with clinical insights, enhancing its utility in assessing heart disease risk.

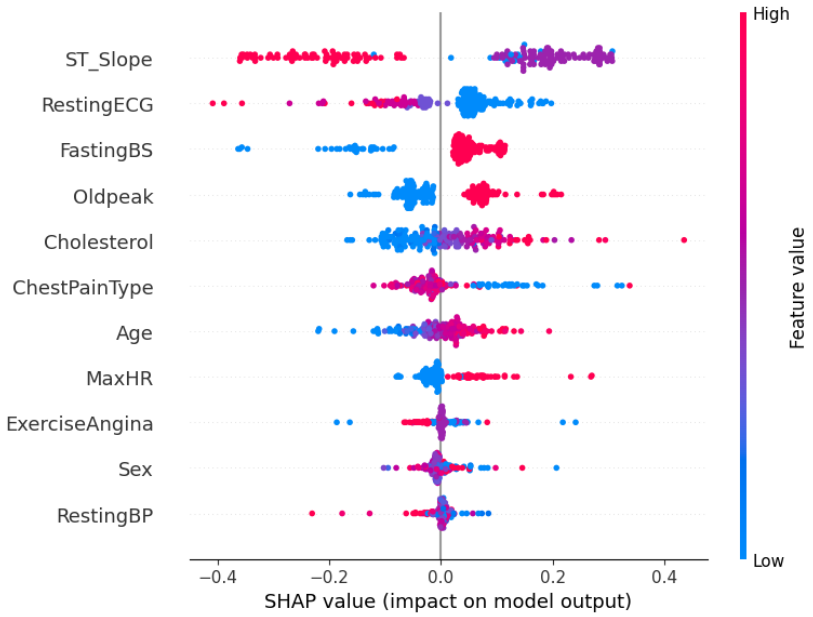


Figure 10: SHAP value summary accross entire dataset

In summary, while MLPs may present challenges in interpretability, leveraging post-hoc explainability methods like SHAP enables us to unravel the model's behavior and enhance its interpretability without compromising its predictive performance. By combining the power of MLPs with the interpretability provided by SHAP, we can make informed decisions and gain valuable insights into complex datasets such as heart failure prediction.Furthermore, our data processing pipeline involved importing, encoding, scaling, and splitting the heart failure dataset to prepare it for subsequent model training and evaluation. These preprocessing steps ensure that the data is appropriately formatted and standardized, enabling effective machine learning analysis and interpretation.

1.4 Neural Additive Models

We applied a Neural Additive Model (NAM) to the dataset. Firstly, we converted the dataframe into Torch tensors, preparing the data for training. Then, we created a custom NAM architecture consisting of multiple FeatureNetworks, each handling one feature. The NAM aggregates the outputs of these networks and applies a sigmoid function to produce the final prediction. During training, we utilized a binary cross-entropy loss function and Adam optimizer. We trained the model for 150 epochs, monitoring the loss at intervals of 10 epochs. The training process achieved a satisfactory accuracy of 82.61% on the test dataset.

The NAM architecture differs conceptually from Logistic Regression and Multi-Layer Perceptrons (MLPs) [NSB23]. While Logistic Regression is a linear model, NAMs are based on non-linear neural networks, allowing them to capture complex interactions between features. However, unlike MLPs, NAMs are structured in a way that each feature is modeled independently by its own network. This design choice makes NAMs more interpretable, akin to Generalized Additive Models (GAMs). GAMs, including NAMs, offer interpretability by decomposing the prediction into contributions from individual features, enabling clear understanding of each feature's importance [MKH+22].

Despite being based on non-linear neural networks, NAMs are more interpretable than MLPs due to their additive structure and feature-wise modeling. In contrast, MLPs operate on the entire feature space simultaneously, making it challenging to understand how each feature contributes to the prediction. Additionally, NAMs provide explicit feature importances, allowing for straightforward interpretation and visualization. Thus, NAMs offer a balance between complexity and interpretability, making them a valuable tool for understanding predictive models in complex datasets.

Additionally we evaluated feature importance by analysing outputs of each feature networks per sample, normalised by the sum of all feature networks outputs predicted by the network. Surprisingly, our analysis output contradicted our SHAP model results, which could be due to the differences in learning strategy of each model (Table 3 and 4). Further investigation would be required in order to better interpret the disparities.

Table 3: Feature Importance Scores Positive samples

Feature	**Sample 1**	**Sample 2**
Age	0.0669	0.0574
RestingBP	0.3109	0.3479
Cholesterol	-0.0129	0.2025
MaxHR	-0.0658	0.0563
Oldpeak	0.2673	0.1845
Sex	0.0244	0.0187
ChestPainType	-0.0399	-0.0306
FastingBS	0.0833	-0.0383
RestingECG	-0.0533	-0.0061
ExerciseAngina	0.0479	0.0367
ST_Slope	-0.0273	-0.0210

1.5 General questions

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

The interpretable/explainable methods, including Logistic Lasso Regression, MLPs with SHAP explanations, and Neural Additive Models (NAMs), consistently identified similar patterns in identifying features important for predicting heart disease outcomes. For example, features like Oldpeak, ST_slope and ExerciseAngina, emerged as significant predictors across these methods, aligning with established medical knowledge.

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.

To convince a doctor to trust the models, we would emphasize their alignment with clinical intuition and established risk factors. For instance, with Logistic Lasso Regression, we would showcase how

Table 4: Feature Importance Scores Negative samples

Feature	**Sample 1**	**Sample 2**
Age	0.0854	0.0305
RestingBP	0.5484	0.3690
Cholesterol	-0.0169	-0.0242
MaxHR	-0.0482	-0.0849
Oldpeak	0.0458	0.0828
Sex	0.0295	0.0317
ChestPainType	-0.0633	-0.2025
FastingBS	0.1008	0.1082
RestingECG	-0.0096	-0.0103
ExerciseAngina	0.0248	0.0267
ST_Slope	0.0273	0.0293

features like Oldpeak, ST_slope and ExerciseAngina reflect known risk factors for heart disease. Similarly, with MLPs and SHAP explanations, we would focus on visualizations illustrating how features contribute to predictions, aligning with medical knowledge. For NAMs, we would highlight their transparency in decomposing predictions into feature contributions, providing explicit feature importances aligned with clinical expectations.

Q3: Elaborate whether the feature importances from the interpretability/explainability methods intuitively make sense to find the respective disease.

Feature importances from the methods intuitively make sense in identifying heart disease risk factors. For example, features like ST_Slope, Oldpeak, and ExerciseAngina consistently emerge as important predictors, aligning with established medical knowledge regarding cardiac health indicators.

Q4: If you had to deploy one of the methods in practice, which one would you choose and why?

For deployment, the choice depends on factors like interpretability and predictive performance. Logistic Lasso Regression is suitable for simplicity and ease of interpretation, while MLPs with SHAP explanations offer deeper insights into feature interactions. NAMs strike a balance between complexity and interpretability, making them optimal for scenarios where both are important. Ultimately, the choice should align with the specific needs of the clinical application.

2 Part 2: Pneumonia Prediction Dataset

In this section, we delve into the classification of chest X-ray images to distinguish between pneumonia and normal cases using deep learning techniques. We explore various interpretability methods to understand model predictions and assess their reliability.

2.1 Exploratory Data Analysis

Upon examining the dataset, we find chest X-ray images sorted into 'Pneumonia' and 'Normal' classes across training ('train'), validation ('val), and testing sets ('test'). While the validation data is evenly distributed, the training set shows a notable class imbalance: it contains 3875 'Pneumonia' cases and 1341 'Normal' cases, whereas the testing set has 390 'Pneumonia' cases and 234 'Normal' cases. This reveals a significant skew in the training dataset, where 'Pneumonia' cases make up around 74.3% of the data (Figure 11). This class distribution skew might introduce bias during model training, potentially causing the model to prioritize features associated with pneumonia over those indicative of normal lung conditions.

Qualitatively examining the dataset through visual inspection of sample images reveals subtle differences between 'Normal' (0) and 'Pneumonia' (1) cases (Figure 12). While some differences may not be immediately apparent, closer examination may reveal characteristic abnormalities such as opacities or infiltrations in the lungs, which are indicative of pneumonia. However, distinguishing between the two classes solely based on visual inspection may be challenging, especially for untrained individuals or when the abnormalities are subtle.

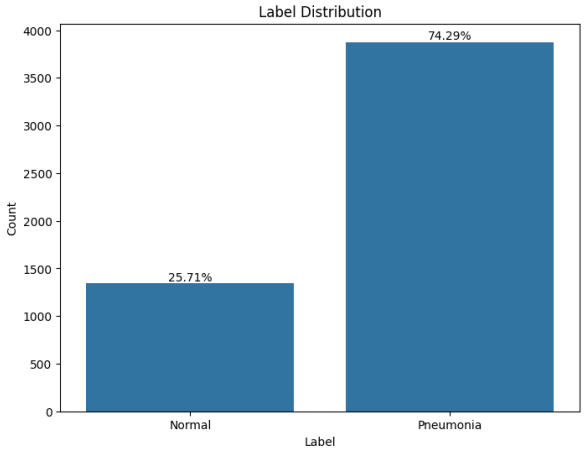
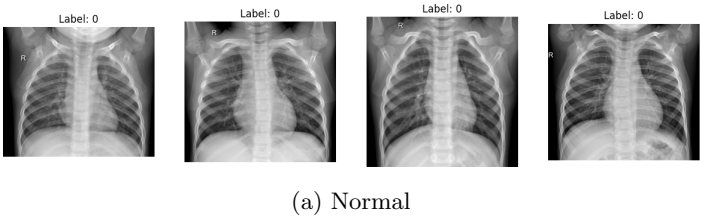
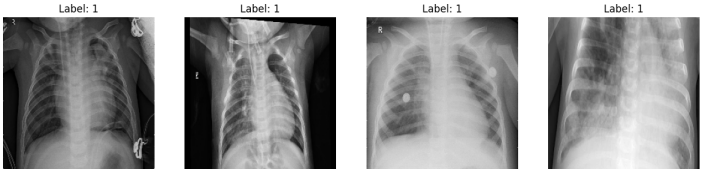


Figure 11: Distribution of Pneumonia cases in training dataset



(a) Normal



(b) Pneumonia

Figure 12: X-ray plot examples on training sets

The data preprocessing steps involve several operations to prepare the training, validation, and testing sets. Firstly, to prevent class imbalance bias during model training, the 'train' and 'val' sets, the number of instances for each 'Normal' and 'Pneumonia' classes is adjusted, ensuring a 50%-50% distribution of both classes in each set. Then, the data is shuffled to randomize the order of instances within each class to prevent the model from learning any order-related patterns in the data that could introduce bias.

The test dataset undergoes resizing to a standard aspect ratio of 224x224 pixels to align with the training and validation datasets. Unlike the training set, no balancing, shuffling, or data augmentation is applied to the test data to ensure unbiased evaluation of model performance. Data generators, constructed using the Keras ImageDataGenerator class, enable efficient batch-wise processing for training, validation, and testing. While data augmentation enhances training set robustness, only rescaling is applied to the validation and test sets, maintaining their integrity for unbiased evaluation.

2.2 CNN Classifier

A convolutional Neural Network (CNN) classifier was designed and trained for the given dataset. The CNN architecture consisted of three convolutional layers followed by max-pooling layers, and two fully connected layers. The ReLU activation function was applied after each convolutional layer, and a dropout layer with a dropout rate of 0.5 was included to prevent overfitting. The model was trained using the Adam optimizer with a learning rate of 0.001 and optimized using the CrossEntropyLoss criterion.

During training, the model was trained on the 'train' dataset, where each epoch involved iterating through the training loader. The training dataset was preprocessed, ensuring a balanced class distribution and shuffling to prevent order-related biases. The model parameters were updated based on the calculated loss, and the optimizer was used to adjust the weights accordingly. This training process was conducted solely on the training dataset.

After training, the performance of the trained model was evaluated on the 'test' dataset using the test function. The test dataset was not subjected to any balancing or shuffling operations to maintain its integrity. The accuracy of the model on the test set was calculated based on the number of correct predictions over the total number of test samples. The reported test accuracy was 0.84.

Additionally, the *predict_class_and_score* function was utilized to predict the class and confidence score of an individual image. This function was applied to a sample image from the training dataset to demonstrate the model's prediction capability and provide insight into its decision-making process. The predicted class for the sample image was 'Pneumonia', with a predicted score of 0.80.

2.3 Integrated Gradients

The integrated gradients method functions as a post-hoc explainability technique, tailored specifically for interpreting image data and overcoming the inherent opacity of Convolutional Neural Networks (CNNs). By generating attribution maps, this method illuminates the crucial regions upon which the CNN relies for its predictions, thereby enhancing interpretability. To implement integrated gradients, we calculate gradients of the model's output with respect to the input image pixels, integrating along a direct trajectory from a baseline image to the target input image. This process provides valuable insights into the contributions of pixels that significantly influence the CNN's decision-making process.

In our implementation, we effectively applied the integrated gradients method to visualize attribution maps for chest X-ray images. These maps adeptly highlight regions crucial for the model's predictions, thereby facilitating a deeper understanding of the CNN's decision-making rationale. When using a generated blurred baseline from a single random image among 'Pneumonia' samples (Figure 13), we observed poor lung targeting of the model on 'Normal' samples. In fact, we noticed stochastic attributions of focus across the images among the 5 'Normal' samples (Figure 14, 16, 18, 20 and 22). However, lung section focus seemed considerably more favorable among the 5 'Pneumonia' samples, although the precision of focus appeared variable between the samples (Figure 15, 17, 19, 21 and 23). This variability may be due to using a unique and general blurred image baseline, which was not customized per each sample.

Conversely, when deploying the default library baseline, the observations on 'Normal' samples revealed more inter-sample consistency in mistargeting the lung area, and alternatively ineffectively focusing on the peripheral rib cages (Figure 24, 26, 28, 30 and 32). Alternatively, the precise lung

area of interest in our 'Pneumonia' sample subset is targeted consistently (Figure 25, 27, 29, 31 and 33). Therefore, we can infer that the baseline selection does seem to have an effect on overall model performance and focus ability on the area of interest. Moreover, the attributions from the use of either baseline seem to be consistent between samples, demonstrating model consistency and robustness.Overall, the integrated gradients method emerges as a powerful tool for interpreting CNNs, offering researchers and practitioners a means to validate and comprehend the model's predictions in image-related tasks.

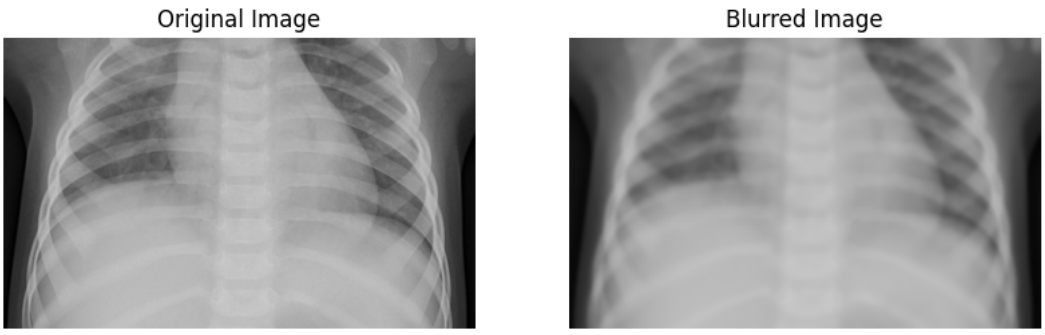


Figure 13: Blurred image to original image comparison

2.4 GRAD-CAM

Grad-CAM, another post-hoc method for generating attribution maps, was implemented to visualize the attribution maps of five healthy and five diseased test samples, similar to the approach taken above. The attribution maps generated by Grad-CAM were examined to assess whether they highlighted sensible regions relevant to the medical conditions under study. Additionally, the consistency of attributions across different samples within each class was evaluated to gauge the reliability of the method.

Consistent with our previous results, when applying our model to the subset 'test' data, we generally observe untargeted focus upon our 'Normal' subset (Figure 34), aligning with the model expectations. The focus is highly variable among samples; however, the primary model focus seems to be mistargeted towards the spine and hips. Conversely, our model succeeded in targeting predominantly the lung area (rib cage area) in the 'Pneumonia' subset, except in 'Pneumonia' sample 5, where the spine and hips were mistargeted (Figure 35). Although mostly targeting the area of interest in all 5 samples, we could observe minor inter-sample variability in the area detection, revealed by the different heat-map intensity gradients between samples.

Comparing with our findings from the previous method (integrated gradients), we note some differences in the attribution maps produced by Grad-CAM. While both methods generally highlight sensible regions relevant to the medical conditions, Grad-CAM seems to exhibit slightly more consistent attributions across samples within each class. However, both methods display some variability in attributions, suggesting that neither method is entirely immune to the inherent complexity of the data and model.

2.5 Data Randomization Test

The data randomization test, introduced in the paper "Sanity Checks for Saliency Maps" [AGM+20], serves as a method to assess the trustworthiness of saliency maps generated by specific attribution methods. Upon retraining the classifier on a training set where labels have been randomly permuted, we obtained a test accuracy of 0.38, which confidently indicates the desired poor ability of the model to distinguish between perturbed and unperturbed samples.

Subsequently, when comparing between the perturbed 'Pneumonia' and unperturbed 'Normal' classifiers on the test samples based on integrated gradients outcome, we could observe that the model equally fails to focus on the expected region in both 'Normal' and 'Pneumonia' samples. This is indicative of favorable model performance given the model learning from randomly permuted labels.

Normal

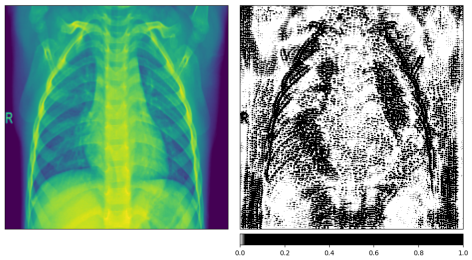


Figure 14: 'Normal' Sample 1

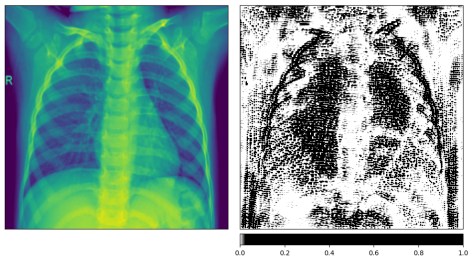


Figure 16: 'Normal' Sample 2

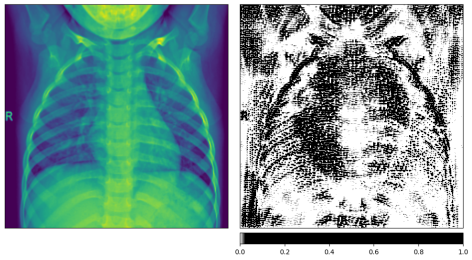


Figure 18: 'Normal' Sample 3

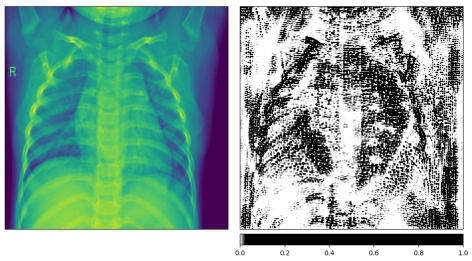


Figure 20: 'Normal' Sample 4

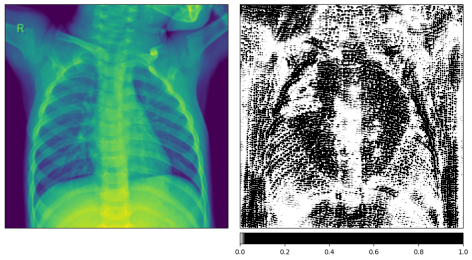


Figure 22: 'Normal' Sample 5

Pneumonia

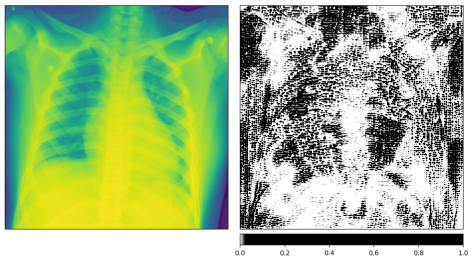


Figure 15: 'Pneumonia' Sample 1

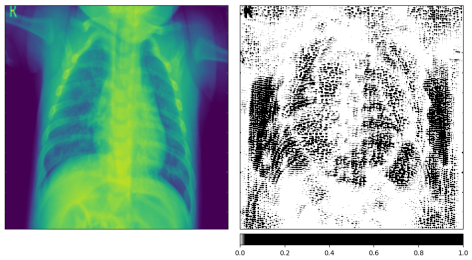


Figure 17: 'Pneumonia' Sample 2

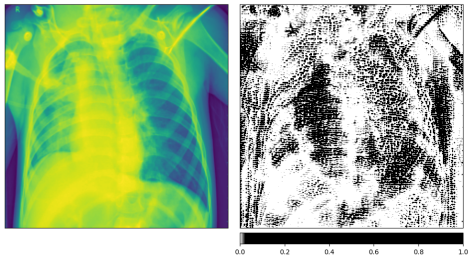


Figure 19: 'Pneumonia' Sample 3

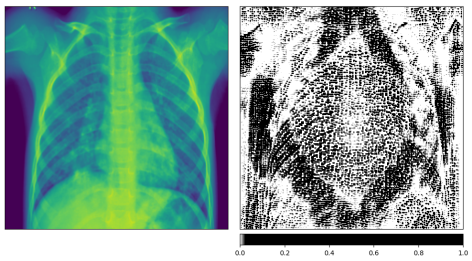


Figure 21: 'Pneumonia' Sample 4

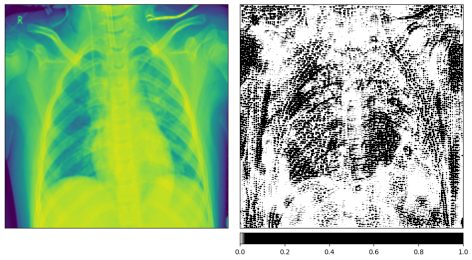


Figure 23: 'Pneumonia' Sample 5

Normal

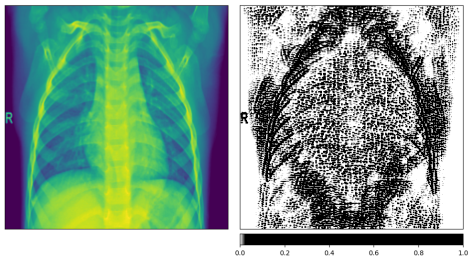


Figure 24: 'Normal' Sample 1

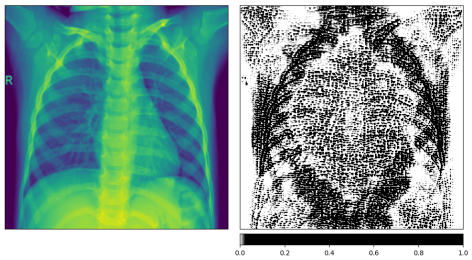


Figure 26: 'Normal' Sample 2

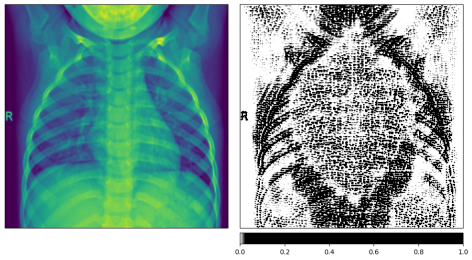


Figure 28: 'Normal' Sample 3

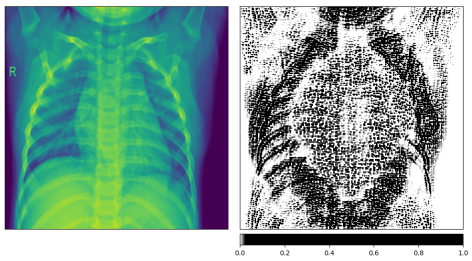


Figure 30: 'Normal' Sample 4

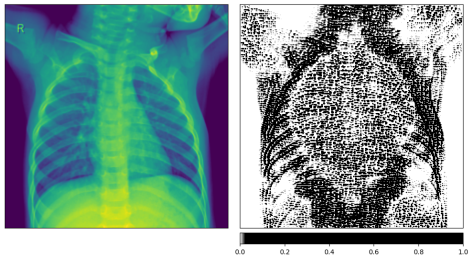


Figure 32: 'Normal' Sample 5

Pneumonia

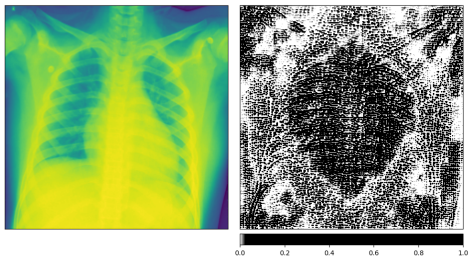


Figure 25: 'Pneumonia' Sample 1

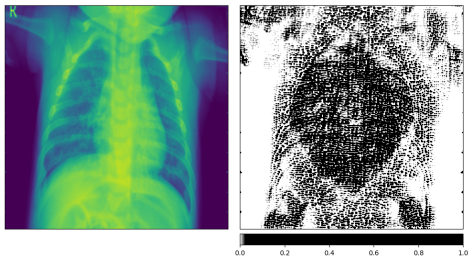


Figure 27: 'Pneumonia' Sample 2

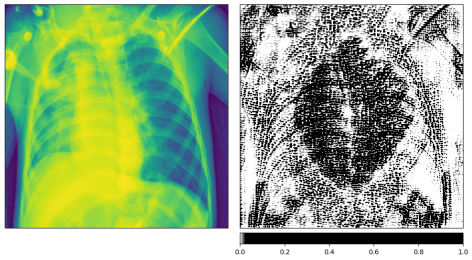


Figure 29: 'Pneumonia' Sample 3

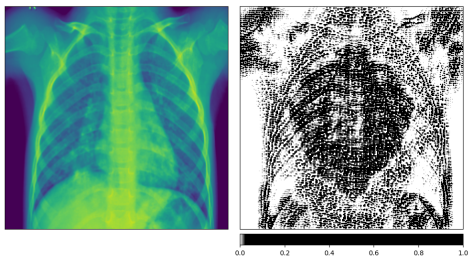


Figure 31: 'Pneumonia' Sample 4

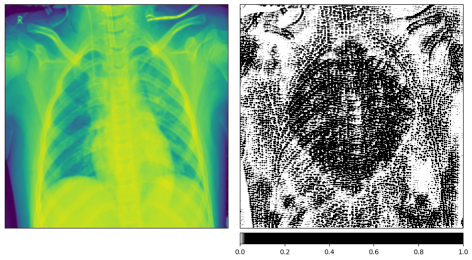


Figure 33: 'Pneumonia' Sample 5

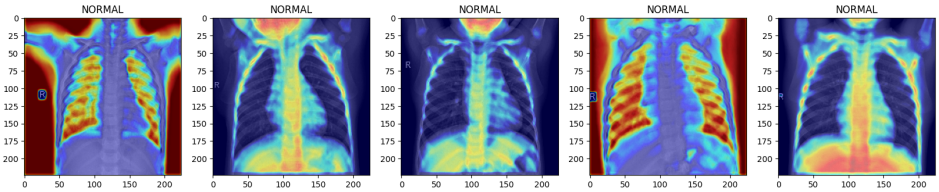


Figure 34: GRAD-CAM 'Normal'

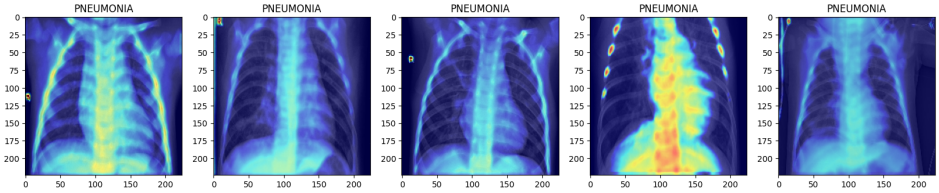


Figure 35: GRAD-CAM 'Pneumonia'

These results not only hold true when deploying our blurred image as baseline (Figure 36-45) but further exacerbate when using the default baseline image (Figure 46-55).

Similarly, when performing and visualizing our Grad-CAM model predictions, we can observe that the model fails to arrive at a consensus area of focus in any of the samples from either class (Figure 57 and 58). This aligns with the findings from the integrated gradients method, indicating that the model's failure to capture meaningful attributions persists even when labels are randomly permuted during training.

Overall, both the integrated gradients and Grad-CAM methods fail the data randomization test, as the attribution maps remain unchanged despite perturbing the labels, suggesting that the attributions captured by these methods may not accurately reflect the relationship between instances and their labels.

2.6 General questions

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

The different interpretable/explainable methods, namely integrated gradients and Grad-CAM, exhibited some consistency in highlighting relevant regions within the chest X-ray images for distinguishing between normal and pneumonia cases. Both methods generally highlighted areas such as the lung region for pneumonia cases, albeit with some variability in the precise focus. However, Grad-CAM seemed to exhibit slightly more consistent attributions across samples within each class compared to integrated gradients.

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.

To convince a doctor to trust the interpretability/explainability results of one of the models, let's consider an example from the integrated gradients method. When examining the attribution maps generated by integrated gradients, we can show the doctor how the method effectively highlights crucial regions in the chest X-ray images that are indicative of pneumonia. By visualizing the areas of focus identified by integrated gradients, such as the opacities or infiltrations in the lungs, we can provide the doctor with valuable insights into the model's decision-making process. Additionally, we can demonstrate the consistency of attributions across different samples, indicating the reliability and robustness of the method in identifying relevant features associated with the disease.

Q3: Elaborate whether the feature importances from the interpretability/explainability methods intuitively make sense to find the respective disease.

The feature importances from the interpretability/explainability methods intuitively make sense to find the respective disease. For instance, in the case of pneumonia, it is known that characteristic abnormalities such as opacities or infiltrations appear in the lungs. Both integrated gradients and Grad-CAM highlight regions in the chest X-ray images that align with these known abnormalities, thereby providing interpretable explanations for the model's predictions. By focusing on these key

Normal

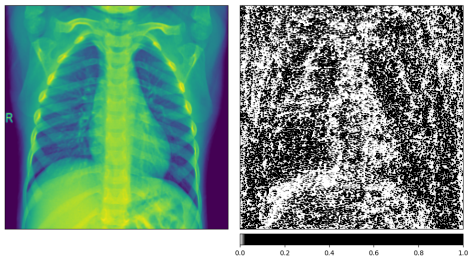


Figure 36: 'Normal' Sample 1

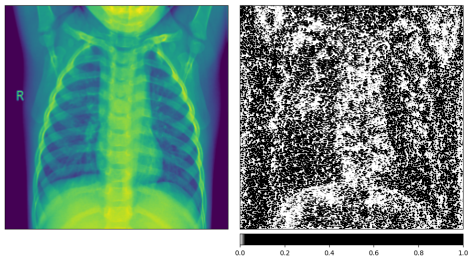


Figure 38: 'Normal' Sample 2

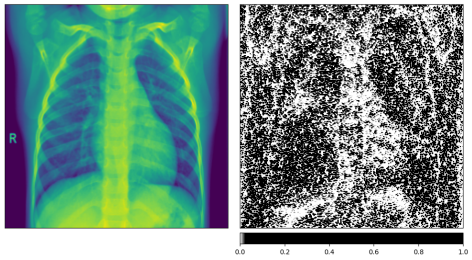


Figure 40: 'Normal' Sample 3

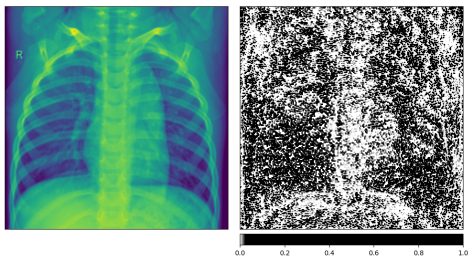


Figure 42: 'Normal' Sample 4

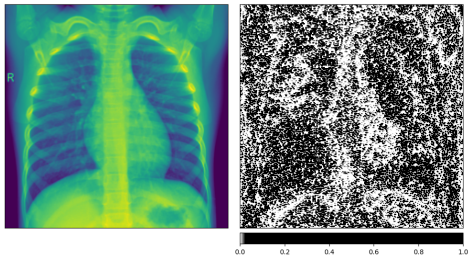


Figure 44: 'Normal' Sample 5

Pneumonia

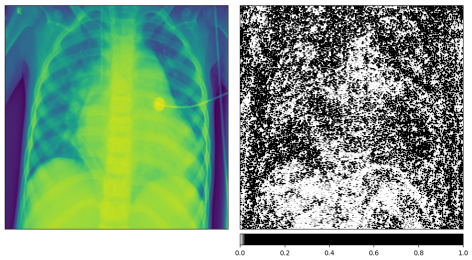


Figure 37: 'Pneumonia' Sample 1

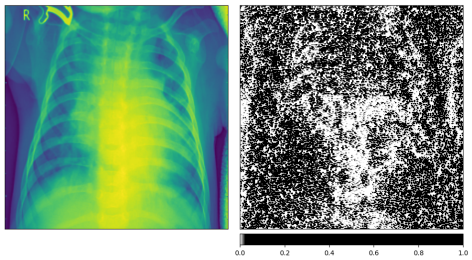


Figure 39: 'Pneumonia' Sample 2

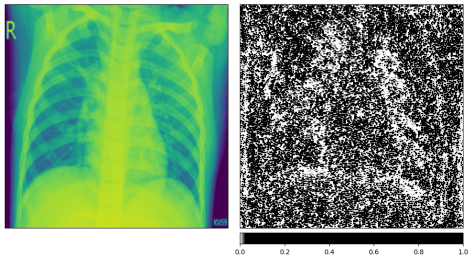


Figure 41: 'Pneumonia' Sample 3

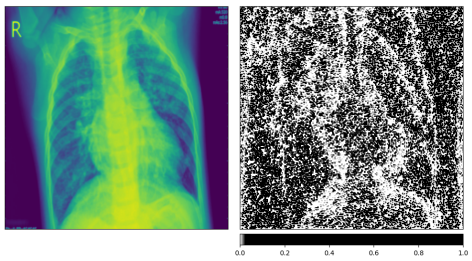


Figure 43: 'Pneumonia' Sample 4

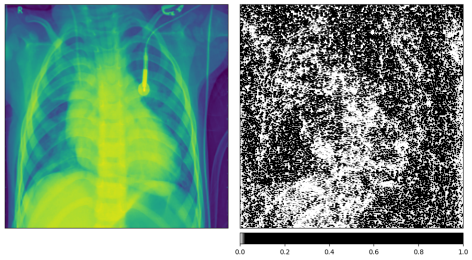


Figure 45: 'Pneumonia' Sample 5

Normal

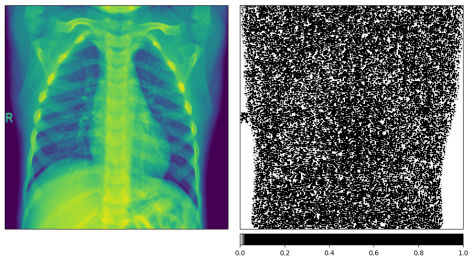


Figure 46: 'Normal' Sample 1

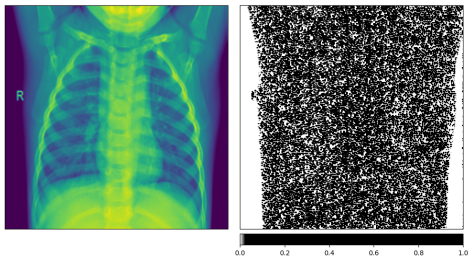


Figure 48: 'Normal' Sample 2

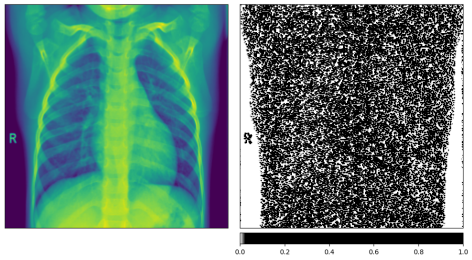


Figure 50: 'Normal' Sample 3

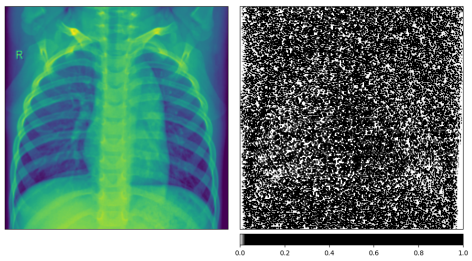


Figure 52: 'Normal' Sample 4

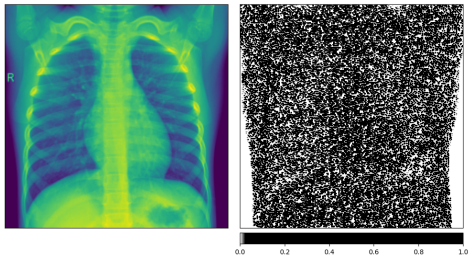


Figure 54: 'Normal' Sample 5

Pneumonia

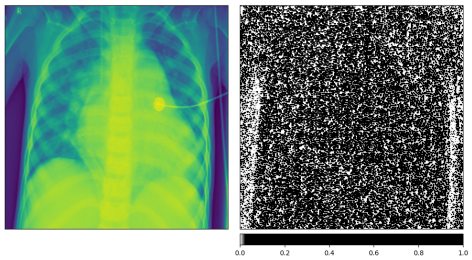


Figure 47: 'Pneumonia' Sample 1

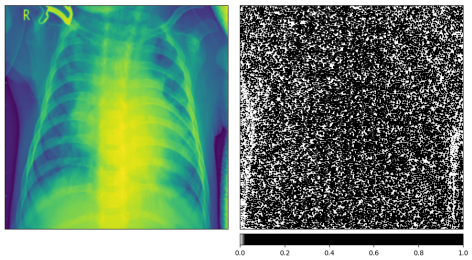


Figure 49: 'Pneumonia' Sample 2

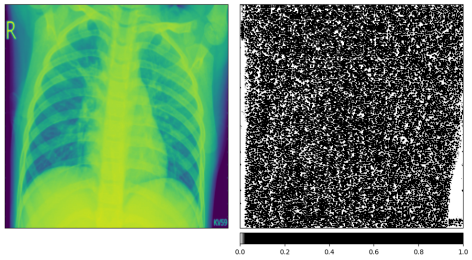


Figure 51: 'Pneumonia' Sample 3

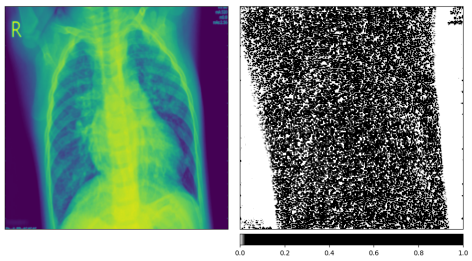


Figure 53: 'Pneumonia' Sample 4

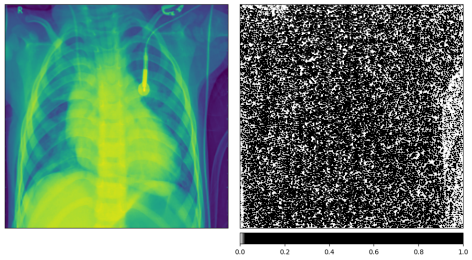


Figure 55: 'Pneumonia' Sample 5

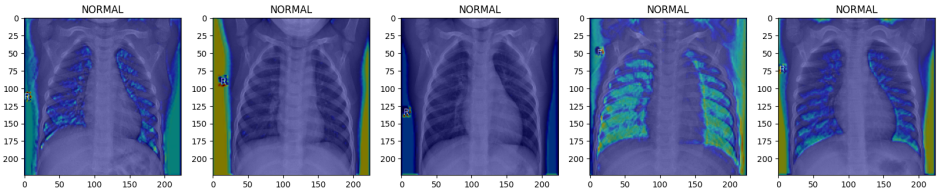


Figure 56: GRAD-CAM 'Normal'

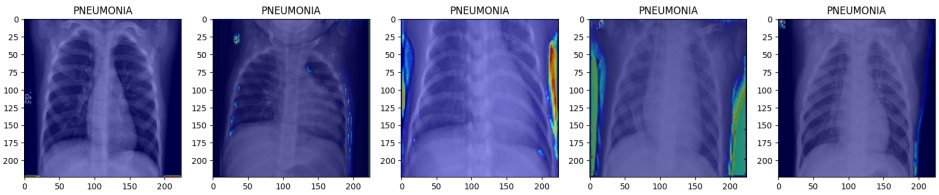


Figure 57: GRAD-CAM 'Pneumonia'

features, the methods effectively capture the underlying characteristics of the disease, enhancing their interpretability and utility in clinical settings.

Q4: If you had to deploy one of the methods in practice, which one would you choose and why?

If we had to deploy one of the methods in practice, we would choose the integrated gradients method. Integrated gradients offer a more detailed and nuanced interpretation of the model's predictions by attributing importance to individual pixels in the input images. This level of granularity allows for a more comprehensive understanding of the model's decision-making process, making it particularly useful for clinical applications where interpretability is crucial. Additionally, the consistency of attributions across samples further enhances the reliability of integrated gradients for guiding clinical decision-making.

References

[AGM+20] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps, November 2020. arXiv:1810.03292 [cs, stat].

[FBM+23] DeLisa Fairweather, Danielle J. Beetler, Nicolas Musigk, Bettina Heidecker, Melissa A. Lyle, Leslie T. Cooper, and Katelyn A. Bruno. Sex and gender differences in myocarditis and dilated cardiomyopathy: An update. *Frontiers in Cardiovascular Medicine*, 10:1129348, March 2023.

[HGS+10] Thomas Huebner, Matthias Goernig, Michael Schuepbach, Ernst Sanz, Roland Pilgram, Andrea Seeck, and Andreas Voss. Electrocardiologic and related methods of non-invasive detection and risk stratification in myocardial ischemia: state of the art and perspectives. *GMS German Medical Science*, 8:Doc27, October 2010.

[KBM24] Anthony H. Kashou, Hajira Basit, and Ahmad Malik. ST Segment. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024.

[LMS+04] G A Lanza, M Mustilli, A Sestito, F Infusino, G A Sgueglia, and F Crea. Diagnostic and prognostic value of ST segment depression limited to the recovery phase of exercise stress test. *Heart*, 90(12):1417–1421, December 2004.

[MKH+22] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 39–68. Springer International Publishing, Cham, 2022.

[noa] Extreme Learning Machine for Multilayer Perceptron | IEEE Journals & Magazine | IEEE Xplore.

[NSB23] J. Naskath, G. Sivakamasundari, and A. Alif Siddiqua Begum. A Study on Different Deep Learning Algorithms Used in Deep Neural Nets: MLP SOM and DBN. *Wireless Personal Communications*, 128(4):2913–2936, February 2023.

[SSG+24] Sophia Sylvester, Merle Sagehorn, Thomas Gruber, Martin Atzmueller, and Benjamin Schöne. SHAP value-based ERP analysis (SHERPA): Increasing the sensitivity of EEG signals with explainable AI methods. *Behavior Research Methods*, March 2024.

[YR01] Salim Yusuf, Srinath Reddy, Stephanie Ôunpuu, and Sonia Anand. Global Burden of Cardiovascular Diseases. *Circulation*, 104(23):2855–2864, December 2001. Publisher: American Heart Association.