# Interpretable and Explainable Classification for Medical Data

**Bondi Francesco**
ETH Zurich
fbondi@ethz.ch
24-942-872

**Fontana Saverio**
ETH Zurich
sfontan@ethz.ch
24-942-971

**Tilman Otto**
ETH Zurich
tiotto@ethz.ch
24-963-308

## 1. Heart Disease Prediction Dataset

### Q1 Exploratory Data Analysis

The dataset includes 918 patient records labeled for heart disease presence, with six numerical features (Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak) and five categorical ones (Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope) (Appendix 4, 5).

Patients with heart disease tend to be older (mean 55.9 vs. 50.6 years), have higher resting blood pressure and cholesterol (with more high-end outliers), lower maximum heart rate, and greater ST depression. Heart disease cases are also more associated with male sex, elevated fasting blood sugar, exercise-induced angina, abnormal ECGs (e.g., ST or LVH), and flat or down-sloping ST segments.

We identified some data quality issues, including physiologically implausible values (e.g., zeros for RestingBP or Cholesterol, negative Oldpeak) and outliers (e.g., Cholesterol $>$ 400 mg/dL). These were treated through imputation and clipping. Class imbalance (55.3% positive) was noted but considered negligible.

For preprocessing, binary features were encoded as 0/1, ST_Slope was ordinal-encoded (Down=0, Flat=1, Up=2), and nominal features (ChestPainType, RestingECG) were one-hot encoded. Missing values (from 0-entries) were imputed using KNNImputer ($k = 5$), Cholesterol outliers capped at the 99th percentile, and negative Oldpeak values set to zero. The dataset was split 80/20 (with stratification), and training features were standardized using StandardScaler.

### Q2 Logistic Lasso Regression

We trained a Lasso-regularized logistic regression model ($\ell_1$ penalty, strength = 1) on the dataset. Crucial preprocessing steps included imputing missing values, one-hot encoding nominal categorical variables, ordinal encoding for ST_Slope, handling outliers, performing a stratified train-test split, and standardizing all features. Standardization ensures coefficient comparability by placing all features on the same scale, which is particularly important for Lasso regularization.

The model achieved an F1-score of 0.886 on the test set. The most influential features with positive weights were Sex (male), FastingBS, and ExerciseAngina, while the strongest negative contributions came from ST_Slope (flat/down-sloping) and certain chest pain types (Atypical, Non-anginal). Moderate contributions included negative weights for MaxHR and ChestPainType_TA, and small positive weights for Oldpeak and Cholesterol. Variables such as Age, RestingECG had near-zero coefficients, and RestingBP was exactly zero (Appendix 6).

While retraining a logistic regression using only the Lasso-selected variables may yield a simpler model, it compromises statistical validity by ignoring the uncertainty from feature selection, invalidating p-values and confidence intervals. Additionally, Lasso may exclude important confounders under penalty, leading to biased estimates. For interpretability, it is more appropriate to rely directly on the sparse Lasso model.

### Q3 Multi-Layer Perceptrons

We trained an MLP with two hidden layers (64, 32 ReLU units, 30 % dropout) and a sigmoid output on the same preprocessed data, using Adam and binary cross-entropy for 30 epochs (batch size 32, 10% validation). The test F1-score was 0.879, comparable to the Lasso model.

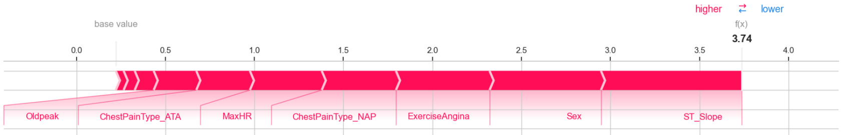


**Figure 1:** Force plot for individual true positive prediction (True Label = 1)

To interpret individual predictions, we used the SHAP `KernelExplainer`. Figure 1 shows a true positive case, while Appendix 8, 9 and 10 display a false negative, a true negative and a false positive using waterfall-style plots. In Figure 1, key features such as a flat `ST_Slope`, `Sex` (male), and `ExerciseAngina` drive the prediction, resulting in a high estimated risk. Conversely, in Figure 8, non-anginal chest pain, an up-sloping `ST_Slope`, and elevated `MaxHR` contribute to the model's misclassification. For the true negative case (Figure 9), `Sex` (female), non-anginal chest pain, and younger `Age` collectively shift the prediction toward a low-risk outcome. However, in the false positive case (Figure 10), a down-sloping `ST_Slope` raises risk, partially offset by `ChestPainType_NAP` and `MaxHR`, yet the model misclassifies.

At the global level, the SHAP summary plot (Appendix 11) identifies `ST_Slope`, `ChestPainType_NAP`, `Sex`, and `ChestPainType_ATA` as the most influential features, closely mirroring the Lasso regression results. High `ST_Slope` values (up-sloping) generally decrease predicted risk, while non-anginal chest pain, being male, and asymptomatic pain increase it. `ExerciseAngina` and `FastingBS` also contribute, though with lower importance.

Comparing local and global SHAP analyses, we observe that while global trends are generally consistent, individual feature contributions can vary substantially across cases. For example, `MaxHR` shows a relatively strong impact in Figures 8 and 10, while `Sex` and `Age` are key in Figure 9. Nonetheless, features like `ST_Slope` consistently emerge as dominant both globally and locally.

### Q4 Neural Additive Models

We trained a Neural Additive Model (NAM) where each feature passes through its own two-layer subnet (32, 16 ReLU units), with outputs summed and passed through a sigmoid. Using Adam (learning rate $10^{-3}$), the model was trained for 50 epochs (batch size 32, 10% validation) and achieved an F1-score of 0.884 on the test set.

Figure 13 (Appendix) shows per-feature shape functions, revealing clinically coherent non-linear effects—e.g., sharp risk increase after age 60, protective high `MaxHR`, and monotonic increases for `FastingBS` and `ExerciseAngina`.

NAMs combine logistic regression's interpretability with neural networks' expressiveness by modeling each feature separately and additively. Compared to logistic regression, they capture richer non-linear effects; unlike MLPs, they avoid feature entanglement, making them more transparent. Each feature's contribution is directly visualizable, offering clear interpretability despite the non-linear structure.

## 2. Pneumonia Prediction

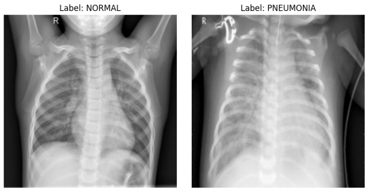


**Figure 2:** X-ray Normal vs Pneumonia

### Q1 Exploratory Data Analysis

The dataset is notably imbalanced, with 3875 pneumonia cases and 1341 normal cases (see 12 in Appendix). Qualitative inspection reveals clear visual differences: Normal scans show well-defined lung fields and visible vascular structures, while pneumonia scans often appear hazy or opaque, suggesting fluid or infiltrates obscuring normal anatomy.

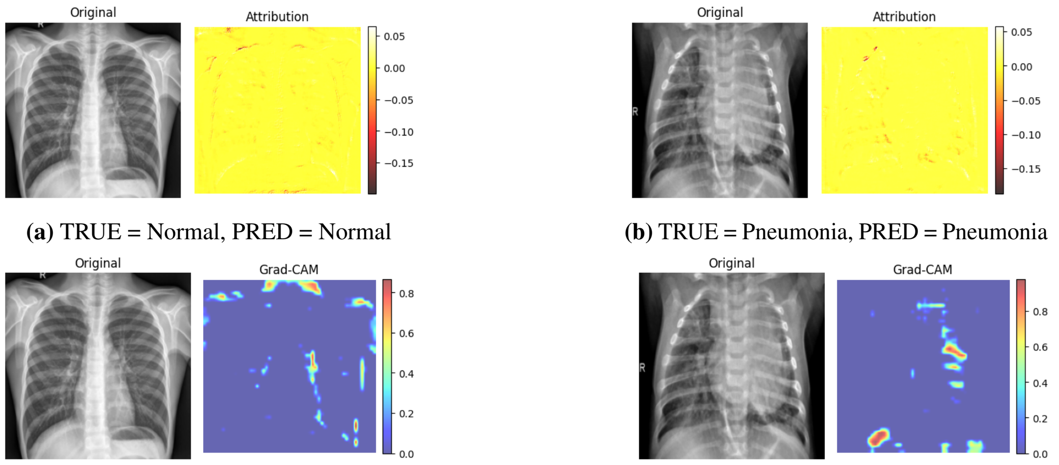
A potential source of bias is scanner-specific artifacts, which may lead the model to learn device-related features rather than true indicators of pneumonia, potentially harming generalization to images from unseen scanners. Additionally, the class imbalance may bias the model towards predicting pneumonia, reducing sensitivity to normal cases. For further analysis, images are resized to $224 \times 224$, converted to grayscale to reduce complexity and emphasize intensity patterns relevant to X-ray interpretation, and scaled to the range $[-1, 1]$ to support stable and efficient model training.

## Q2 CNN Classifier

Our CNN classifier consists of a feature extractor and a classifier. The feature extractor uses three Conv2D layers (32 to 128 filters), each with ReLU and 2×2 max pooling. The flattened features pass through a 256-neuron dense layer, dropout (0.5), and a single-neuron output for binary classification. Test performance: AUROC=0.905, AUPRC=0.913, F1=0.838.

## Q3 Integrated Gradients

We applied Integrated Gradients to five samples each from pneumonia and normal lung scans. The resulting attribution maps (Figure 3) use heatmaps where brighter regions indicate stronger contributions toward the pneumonia classification, and darker regions support normal predictions. In pneumonia cases, high-attribution areas corresponded to clinically relevant features such as localized opacities or infiltrates, though their positions varied across the lungs. Normal cases exhibited more diffuse attributions, often along lung boundaries and clear regions, with some emphasis on stable anatomical structures like central lung zones. A minor artifact was observed where the letter *'R'*, occasionally present in X-rays, received attribution despite being clinically irrelevant (Appendix 14b). To reduce attribution noise, we also tested gray, random noise, and blurred baselines, but these alternatives provided no clear improvement over the black baseline and often reduced interpretability (visualizations in Appendix 15).



**(a)** TRUE = Normal, PRED = Normal

**(b)** TRUE = Pneumonia, PRED = Pneumonia

**(c)** TRUE = Normal, PRED = Normal

**(d)** TRUE = Pneumonia, PRED = Pneumonia

**Figure 3:** IG and Grad-CAM attribution maps for two patients

## Q4 Grad-CAM

We also applied Grad-CAM to visualize class-discriminative regions, where warm colors (red/yellow) indicate strong predictive relevance and cool colors (blue) denote low impact. In pneumonia cases, Grad-CAM consistently highlighted patchy opacities across different lung regions, aligning well with known clinical indicators. Normal samples showed more diffuse activations or focus around anatomical landmarks, suggesting reliance on broader healthy features. Visual inspection of Figure 3 indicates that Grad-CAM more effectively emphasized diagnostically relevant regions and largely ignored irrelevant artifacts such as the letter "R" (see Appendix 14d). Overall, Grad-CAM provided more focused and clinically coherent attribution compared to other methods.

### Q5 Data Randomization Test

To assess the reliability of saliency methods, we applied the data randomization test proposed by Adebayo et al.[1]by training a CNN on data with randomized labels. As evaluation metrics, we computed the Spearman rank correlation for both absolute and raw saliency values.[2]

For Integrated Gradients, the average Spearman correlations between correctly and randomly labeled models were low (Mean Spearman (ABS): 0.2560; Mean Spearman (Diverging): 0.0214), indicating strong dependence on label structure and suggesting that the method passes the data randomization test. In contrast, Grad-CAM yielded a higher average Spearman correlation of 0.3258,[3] implying some label-independent consistency and a partial failure of the test.

Figures 16 and 17 (Appendix) compare attribution maps (from both methods) for four representative patients under both correct and randomized labeling, illustrating the differing sensitivities. Baseline comparisons using random noise showed near-zero Spearman correlations (e.g., -0.0001 for Integrated Gradients, -0.0002 for Grad-CAM), confirming the meaningfulness of the above results.

## 3. General Questions

### Q1 Consistency of Interpretability Methods

**Part 1:** All three approaches identify ST segment slope and chest pain type as top predictors, alongside sex and exercise-induced angina. However, the linear Lasso coefficients capture only straight-line effects, whereas SHAP and NAM reveal non-linear patterns (e.g. risk accelerating after age 60 or plateauing at high MaxHR). This reflects the richer expressiveness of the neural methods. **Part 2:** Both Integrated Gradients and Grad-CAM often highlighted lung regions with opacities or anomalies. However, their exact attribution maps varied in sharpness and focus. Grad-CAM tended to be coarser but more localized, while Integrated Gradients produced smoother but broader relevance regions. Overall, they identified overlapping regions, suggesting partial consistency.

### Q2 Convincing a Clinician

**Part 1:** With the NAM's shape plots, age rising sharply after 60, MaxHR sloping downward, and ST_Slope peaking then falling, we can overlay a patient's values to show exactly how each factor drives the final risk (F1 = 0.884), offering a transparent, clinically intuitive explanation. **Part 2:** With Grad-CAM visualizations, we can overlay the attribution maps directly on chest X-rays, showing the regions the model focused on. If these align with what radiologists expect in pneumonia, this builds confidence that the model mimics medical reasoning rather than exploiting spurious patterns.

### Q3 Clinical Plausibility of Feature Importances

**Part 1:** Key predictors above correspond to established risk factors: ST slope anomalies and atypical chest pain reflect myocardial ischemia, male sex and exertional angina indicate higher baseline risk, and metabolic measures (fasting glucose) contribute additively. The non-linear shapes (e.g. U-shaped age) further mirror physiological processes, confirming that model importances are clinically sensible. **Part 2:** Mostly yes. The attribution maps highlight regions in the lungs where pneumonia typically manifests, like peripheral opacities. However, occasional off-target attributions suggest some risk of dataset bias or overfitting to background artifacts.

### Q4 Method Choice for Deployment

**Part 1:** We would deploy the Neural Additive Model. It achieves top predictive performance (F1 = 0.884) while providing direct, per-feature shape plots that clinicians can inspect and validate, offering "glass-box" transparency without sacrificing non-linear expressiveness. **Part 2:** We would deploy the Grad-CAM-enhanced CNN, as it provides intuitive visual explanations that clinicians can validate directly against known radiological features. Despite being coarser than Integrated Gradients, its alignment with spatial pathology is more interpretable in practice.

[1] "Sanity checks for saliency maps." Advances in Neural Information Processing Systems 31 (2018).

[2] SSIM and HOG were not implemented, as they are tailored for natural images and unsuitable for X-ray data.

[3] We only report one correlation for Grad-CAM due to the ReLU-induced non-negative nature of its heatmaps.

# Appendix

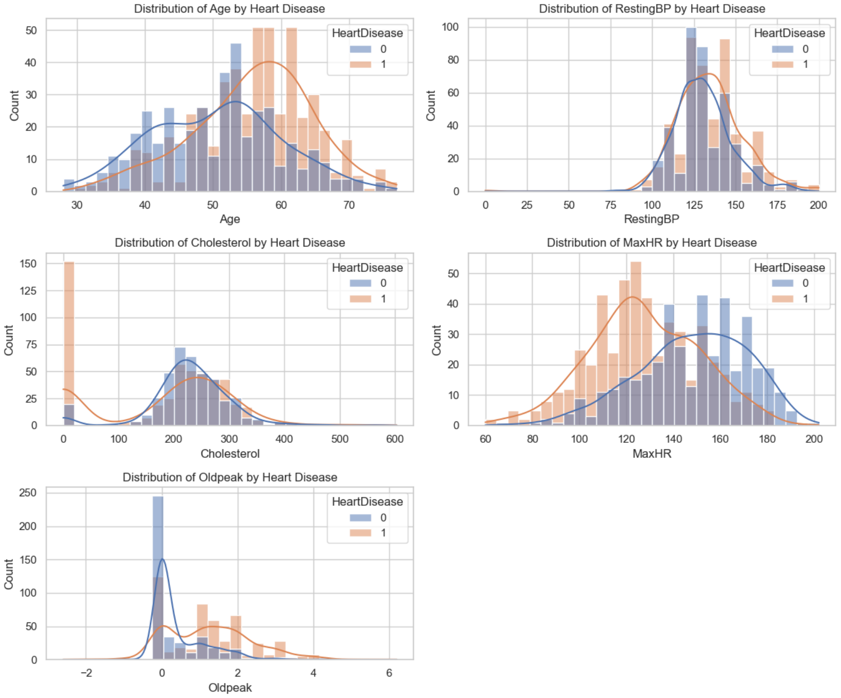


**Figure 4:** Part 1 - Data distribution for continuous variables

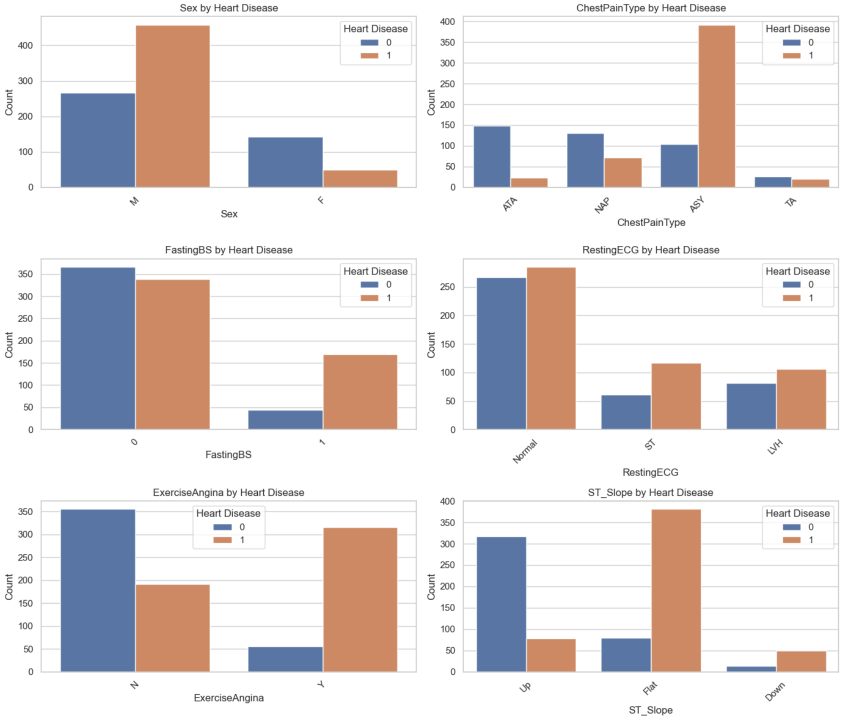


**Figure 5:** Part 1 - Data distribution for categorical variabes

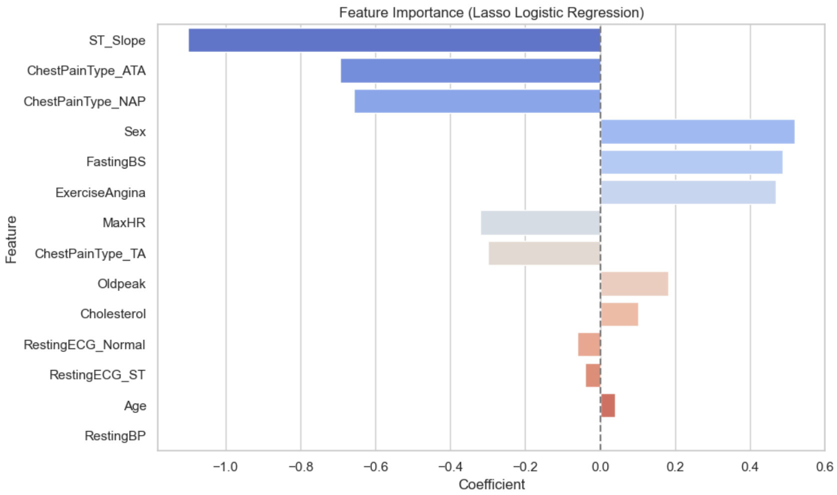


**Figure 6:** Part 1 - Feature importance as given by the coefficients of the Lasso logistic regression.

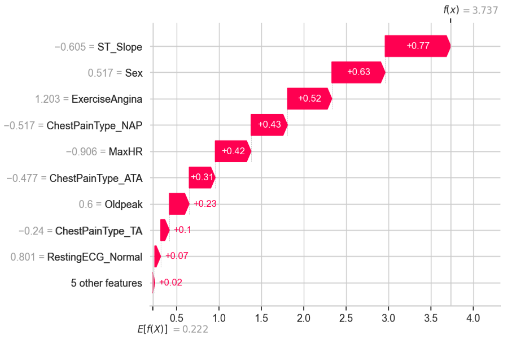


**Figure 7:** Part 1 - SHAP explanation for a true positive prediction (model output $f(x) = 3.737$).

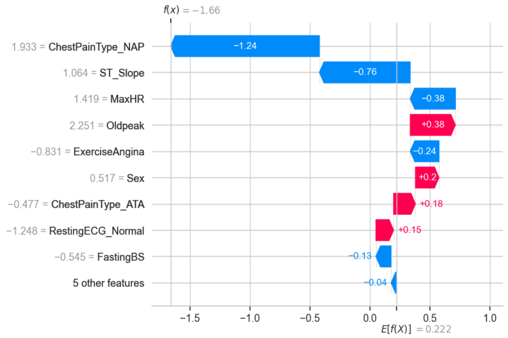


**Figure 8:** Part 1 - SHAP explanation for a false negative prediction (model output $f(x) = -1.66$).

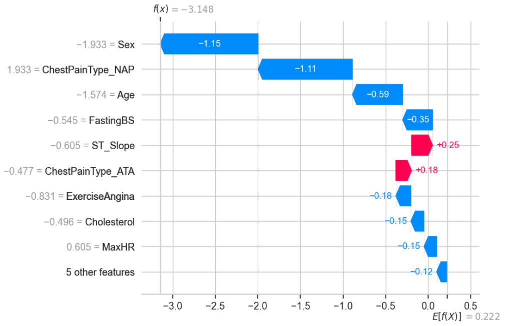


**Figure 9:** Part 1 - SHAP explanation for a true negative prediction (model output $f(x) = -3.148$).

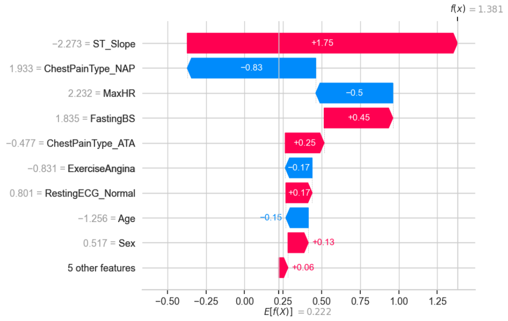


**Figure 10:** Part 1 - SHAP explanation for a false positive prediction (model output $f(x) = 1.381$).

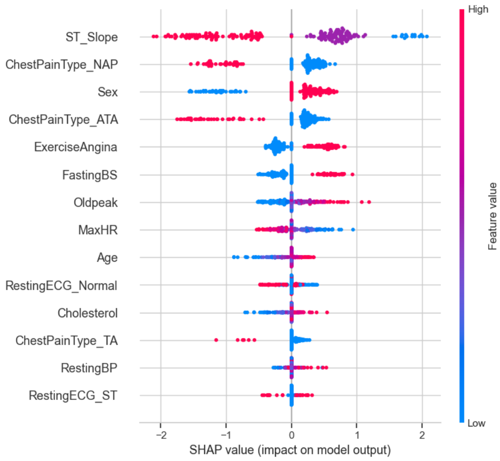


**Figure 11:** Part 1 - Global feature importance from SHAP values across all test samples.

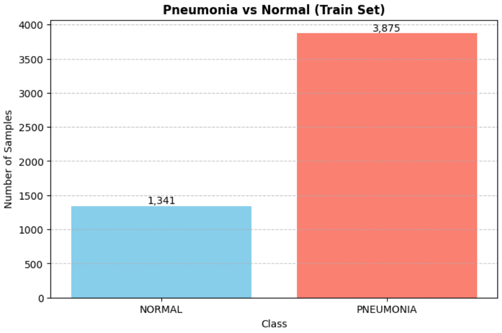


**Figure 12:** Part 2 - Label distribution for Pneumonia prediction

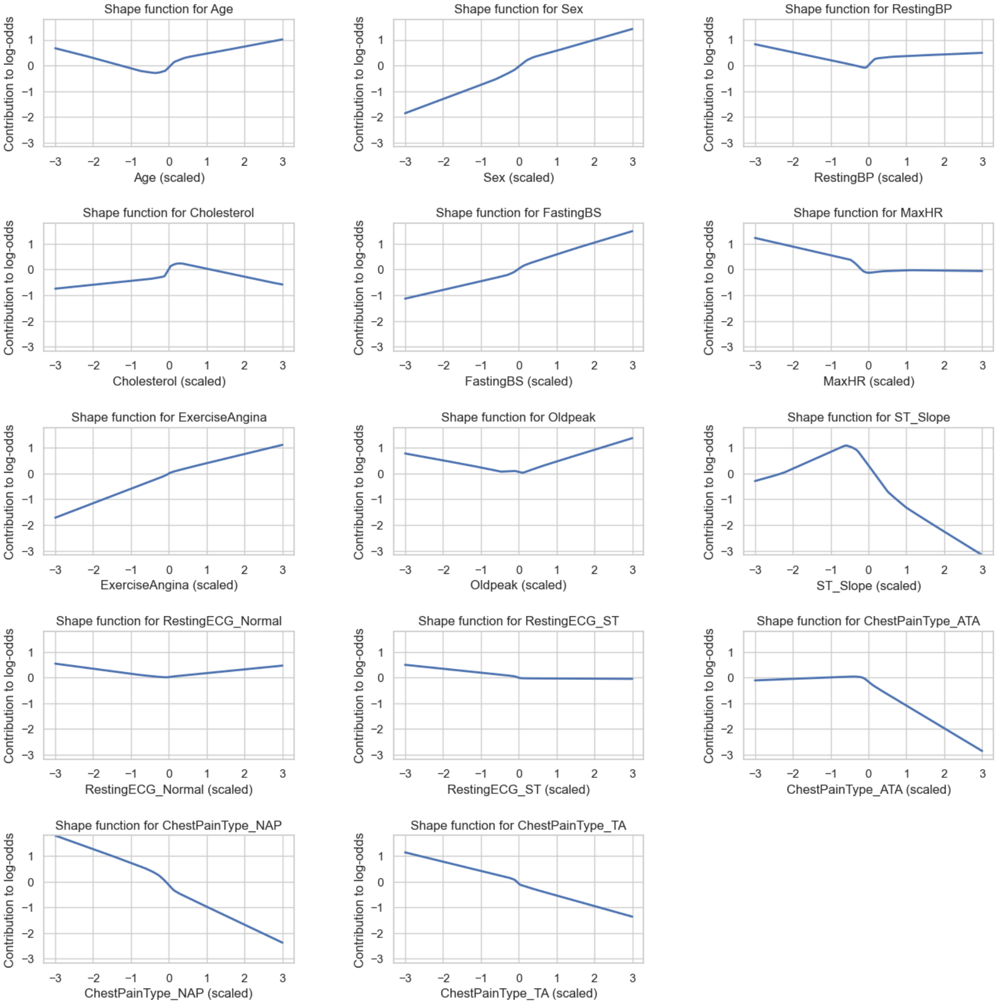
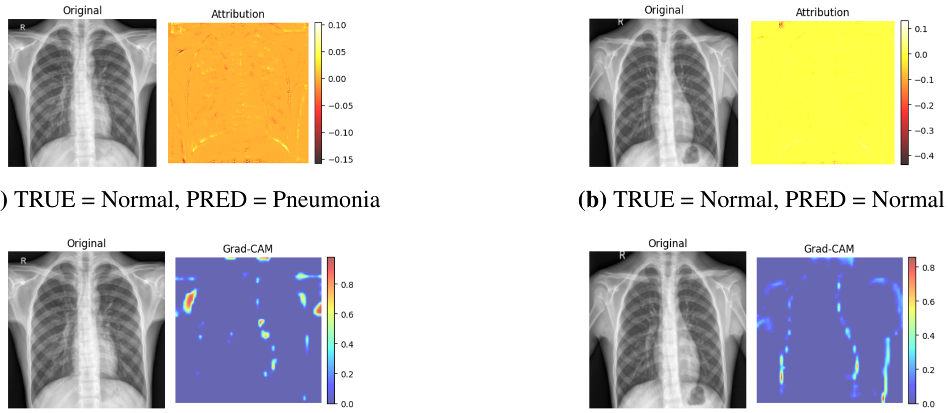


**Figure 13:** Part 1 - Shape functions for each feature in the NAM, showing contribution to log-odds versus standardized input.



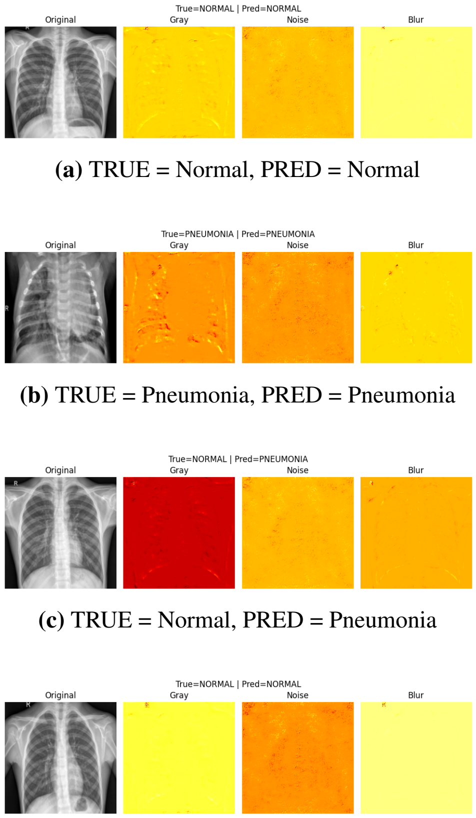
**(a)** TRUE = Normal, PRED = Pneumonia

**(b)** TRUE = Normal, PRED = Normal

**(c)** TRUE = Normal, PRED = Pneumonia

**(d)** TRUE = Normal, PRED = Normal

**Figure 14:** Part 2 - IG and Grad-CAM attribution maps for two patients


True=NORMAL | Pred=NORMAL
Original
Gray
Noise
Blur
True=PNEUMONIA | Pred=PNEUMONIA
Original
Gray
Noise
Blur
True=NORMAL | Pred=PNEUMONIA
Original
Gray
Noise
Blur
True=NORMAL | Pred=NORMAL
Original
Gray
Noise
Blur

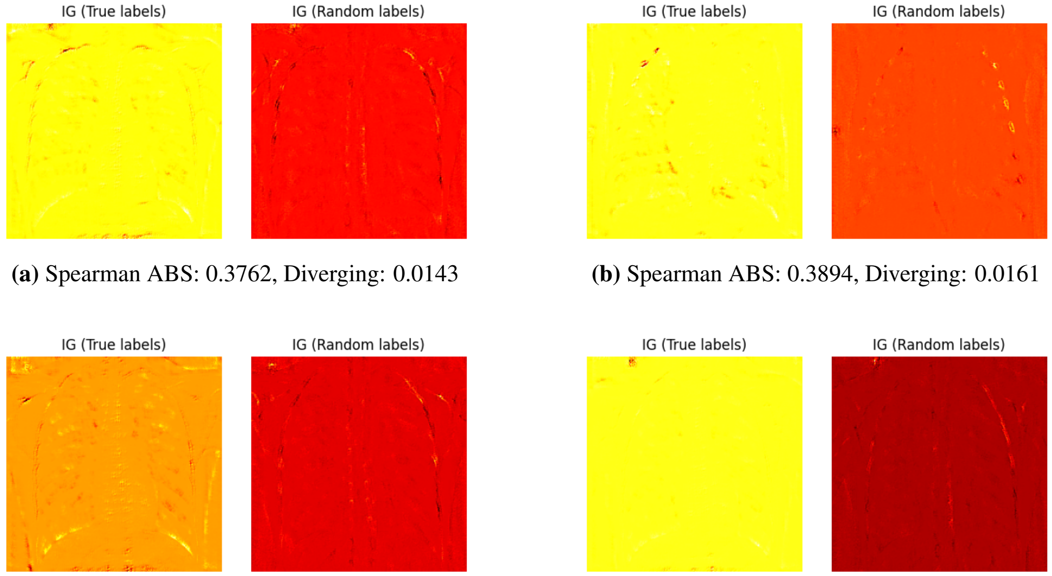
**(a)** TRUE = Normal, PRED = Normal

**(b)** TRUE = Pneumonia, PRED = Pneumonia

**(c)** TRUE = Normal, PRED = Pneumonia

**(d)** TRUE = Normal, PRED = Normal

**Figure 15:** Part 2 - IG attribution maps with different baselines: gray, random noise, and gaussian blurred



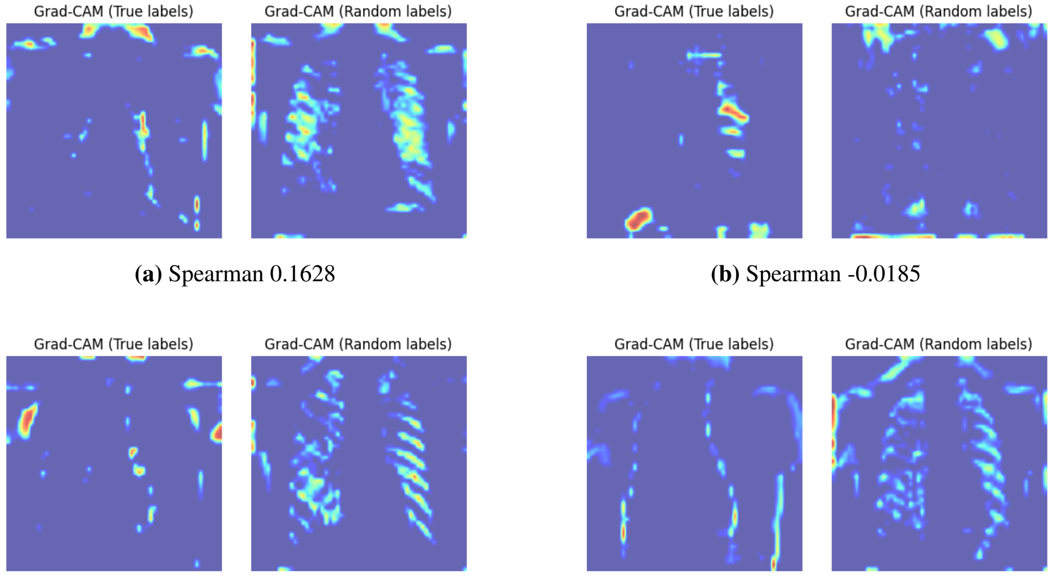
**(a)** Spearman ABS: 0.3762, Diverging: 0.0143

**(b)** Spearman ABS: 0.3894, Diverging: 0.0161

**(c)** Spearman ABS: 0.3386, Diverging: 0.0051

**(d)** Spearman ABS: 0.3762, Diverging: 0.0143

**Figure 16:** Part 2 - IG attribution with data randomization test



**(a)** Spearman 0.1628

**(b)** Spearman -0.0185

**(c)** Spearman -0.1229

**(d)** Spearman 0.0348

**Figure 17:** Part 2 - Grad-CAM attribution with data randomization test