Interpretable and Explainable Classification for Medical Data

Ángel Labordet; (22-908-693) alabordet@ethz.ch
Jolanda Malamud; (12-921-938) jmalamu@ethz.ch

Code: https://github.com/jolandamalamud/ml_healthcare_project2

1 Part 1. Heart Disease Prediction Dataset

1.1 Q1 – Exploratory Data Analysis

Q1.1 Feature & label overview. The *train–val* split contains **734** patients of whom **398** (54 %) are positive. Apart from a spike at *RestingBP*=0, zeros in *Cholesterol*, and nine negative *Oldpeak* values, numeric variables are well behaved. Categorical levels are balanced; the rarest is `ChestPainType=TA` (4 %).

Q1.2 Pitfalls and fixes. Values that are physiologically impossible (BP=0, Chol=0, Oldpeak<0) are turned into `NaN` and KNN-imputed ($k = 5$) in z-score space. The mild 54/46 class imbalance is kept in check by stratified CV and `class_weight=balanced`. Full details appear in Appendix A.1.

Q1.3 Pre-processing blueprint (used everywhere). **Numeric branch:** `StandardScaler` → `KNNImputer` → floor at clinical minima → `StandardScaler`. **Categorical branch:** mode-impute → `OneHotEncoder(drop='first')`. The un-fitted transformer is saved as `preprocessor.pkl` and cloned inside every CV fold, preventing leakage (Appendix A.2).

1.2 Q2 – Logistic Lasso Regression

Q2.1 Model. We wrap `preprocessor.pkl` in a `Pipeline` with `LogisticRegression(penalty=L1, solver=saga, class_weight=balanced)`. A 5-fold grid search over $C \in 10^{[-3,\dots,2]}$ selects $\hat{C} = 0.53$.

Q2.2 Why the coefficients are comparable. *Double standardisation* of numerics (before & after imputation) gives each column SD 1; one-hot dummies are 0/1. Thus the L1 penalty treats every feature on an equal footing and the magnitudes of β carry meaning. **Note:** ideally we would apply a *group-lasso* so that all dummies derived from the same categorical are kept or dropped together; the packages tested (`group_lasso`, `groupyr`, `lightning`) failed in our environment, so we interpret isolated zeroed dummies with caution.

Q2.3 Performance. Outer-CV: balanced-accuracy 0.867 ± 0.047, macro-F1 0.867 ± 0.048.

Q2.4 What the model learned. Appendix Fig. A.2 plots the 15 strongest coefficients. ST-slope, chest-pain type and sex dominate; *Oldpeak* and exercise-induced angina also push predictions, while two ECG dummies are shrunk to zero.

Q2.5 Is it wise to refit an un-penalised logistic on the kept features? Usually not. The refit ignores that feature selection already used the data, so its CIs are far too narrow (post-selection bias). Nested CV or post-selection inference would be required; otherwise, the single regularised model is the sound choice.

1.3 Q3 – Multi-Layer Perceptron (MLP) + SHAP

Q3.1 Model. The frozen `preprocessor.pkl` feeds a one-hidden-layer MLP tuned with `BayesSearchCV` (80 trials, 5-fold CV, balanced-accuracy). The search yields `hidden_layer_sizes=(66)`, `activation=relu`, `alpha`=7.6×10^{-6}, `lr_init=0.01`, `solver=adam`. *Outer-CV*: Bal. Acc $= 0.877$. *Held-out test set*: Bal. Acc $= 0.821$, Macro-F1 $= 0.824$.

Q3.2 SHAP explanations. Kernel-SHAP (background = all 734 train–val rows) decomposes the log-odds for each of the 184 test patients. *Local view*: four representative waterfalls (two positives, two negatives) are shown in Appendix Fig. A.3. *Global view*: Appendix Fig. A.3 combines (a) mean |SHAP| bar-plot, (b) layered violin (SHAP vs. feature value), and (c) heat-map of patients × top-10 features.

Q3.3 Consistency of importances. Across local waterfalls and global plots the same drivers recur; ST-slope dummies, chest-pain dummies, *Oldpeak*, *Sex_M*, *Cholesterol*. Ranking can shuffle patient-to-patient due to threshold effects (e.g. extreme *Oldpeak*) and correlated dummies, yet the *set* of key variables remains stable, mirroring the Logistic-Lasso list and boosting confidence in the model's explanations.

1.4 Q4 – Neural Additive Model (NAM)

Q4.1 Model. Each input feature drives its own two-layer ReLU subnet (g_i); the resulting logits $\sum_i g_i(x_i)$ are passed through a sigmoid.A Bayesian hyper-parameter search (60 trials, 5-fold stratified CV,balanced-accuracy objective) selected `num_units`=110, `dropout`=0, and learning rate $\eta = 5 \times 10^{-4}$. *Cross-validation:* balanced-accuracy 0.877 ± 0.046. *Held-out test:* balanced-accuracy 0.806, macro-F1 0.807. The final leak-proof pipeline is stored as `nam_pipe.pkl`.

Q4.2 Feature importance. Because the prediction is $\sigma\big(\sum_i g_i(x_i)\big)$, the signed contribution $g_i(x_i)$ is available *per sample*. We plot the mean $|g_i|$ over all 184 test patients (AppendixFig. A.4); ST-slope dummies, chest-pain type, sex and cholesterol emerge as the dominant factors, echoing the Logistic-Lasso and SHAP findings [1].

Q4.3 Conceptual comparison.

	Logistic-Lasso	MLP	NAM
Form	linear	universal NN	*additive* NN
Capacity	low (bias ↑)	high (variance ↑)	mid-ground
Interactions	none	all orders	none (by design)
Interpretability	global β's	post-hoc only	intrinsic 1-D shapes
Test macro-F1	0.867	0.824	0.807 [**?**]

Q4.4 Why NAMs are (much) more interpretable than MLPs. A NAM enforces an **additive** logit $\sum_i g_i(x_i)$; every $g_i : \mathbb{R} \to \mathbb{R}$ is a smooth, one-dimensional curve. Hence:

(a) *Exact local explanations:* $g_i(x_i)$ is the signed share of feature i in *that* prediction.

(b) *Global explanations:* one 1-D plot per feature fully describes the model; no entangled weights to decipher.

(c) *Monotonicity / shape constraints* can be imposed directly on g_i, impossible in a dense MLP.

An ordinary MLP mixes all features in high-dimensional hidden layers, so neither per-feature contributions nor global response shapes are recoverable without approximate post-hoc tools, making it far less transparent [1, 2].

2 Part 2: Pneumonia Prediction Dataset

2.1 Q1: Exploratory Data Analysis

Q1.1 Dataset. The chest X-ray dataset comprises 5216 images, with a pronounced class imbalance: 3875 (74.3%) pneumonia and 1341 (25.7%) normal cases. Of the 3875 pneumonia cases, 2530 (65. 3%) are bacterial and 1345 (34.7%) are viral.

Q1.2 Visual Analysis. Exploratory analysis of sample images reveals distinct visual differences (cf. Fig 1): normal X-rays show radiolucent (transparent/black) lung fields with clear heart borders and blood vessels, whereas pneumonia cases show white spots and areas in the darker background of the lungs.

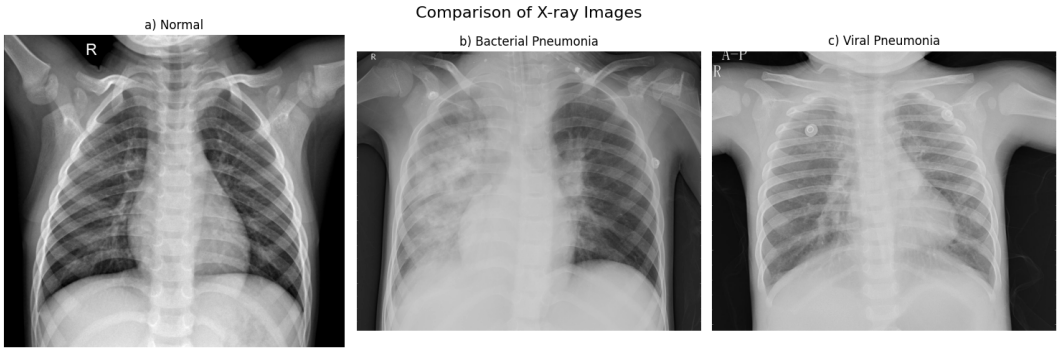


Figure 1: Visual analysis of sample chest X-ray images. a) Normal lungs displaying clear, uniformly dark lung fields with well-defined heart borders. b) Bacterial pneumonia showing dense, localized unilateral opacities. c) Viral pneumonia demonstrating more diffuse, bilateral, and less dense interstitial opacities spread throughout the lungs.

Q1.3 Class Imbalance. A potential bias in model learning is the class imbalance in the training dataset, where normal cases are underrepresented compared to pneumonia cases. This could cause the model to favor majority class predictions, reducing its sensitivity to detect actual pneumonia. We address this imbalance by using class weighting during training.

Q1.4 Preprocessing. We perform the following preprocessing steps: i) resizing to 224×224 pixels, ii) normalization, and iii) data augmentation, i.e. horizontal flips and rotations.

2.2 Q2: CNN Classifier

Q2.1 CNN Design. We designed a deep CNN that processes medical images through 4 convolutional blocks with increasing filter complexity, batch normalization, ReLU activation, and max pooling, before classifying them as normal or pneumonia using fully connected layers with dropout regularization and a weighted cross-entropy loss function to handle class imbalance.

Q2.2 CNN Performance. The CNN achieved strong performance on the pneumonia classification task, with an accuracy of 0.85, a recall of 0.95, precision of 0.84, and an F1 score of 0.89 (cf. confusion matrix in Appendix Fig. A.5).

2.3 Q3: Integrated Gradients

Q3.1 Implementation. We implemented the Integrated Gradients method, which quantifies feature importance by measuring gradients of model predictions along a linear interpolation path from a reference baseline to the input X-ray image. This approach accumulates attributions to identify regions most critical to classification decisions. We visualized the overlap of attribution map and X-ray image for 5 normal and 5 pneumonia samples (cf. Appendix Fig. A.7).

Q3.2 Attribution Maps. The Integrated Gradients attribution maps highlight different regions in pneumonia and normal cases. In pneumonia cases, the maps emphasize affected lung areas (cf. example case in Fig. A.6 (a)). In normal cases, the attribution is distributed across the skeletal structures, the diaphragmatic contours, and the heart boundary (cf. example case in Appendix Fig. A.6 (b)).

Q3.3 Consistency. The attribution maps demonstrate high consistency across samples, with reproducible patterns of feature importance for both pneumonia and normal cases.

Q3.4 Baseline Input. The choice of baseline input significantly influences attribution map. After evaluation of multiple baseline options (black, white, grey, blurred, and average image), we selected the black (zero) baseline as optimal for our application, as it provides superior visualization of lung structures and pathological features.

2.4 Q4: Grad-CAM

Q4.1 Implementation. We implemented the Grad-CAM method, which uses the gradients of a target prediction flowing into the final convolutional layer (here layer 4) to create a heatmap that highlights the regions of the input image most influential to the model's decision. We visualized the overlap of Grad-CAM map and X-ray image for 5 normal and 5 pneumonia samples (cf. Appendix Fig. A.9).

Q4.2 Attribution Maps. In pneumonia cases, Grad-CAM consistently emphasizes areas of abnormal tissue, often localized in the lower lobes, reflecting radiological patterns such as consolidation (cf. example case in Appendix Fig. A.8 (a)). In contrast, normal cases exhibit diffuse activation over cardiac silhouettes, and diaphragmatic contours (cf. example case in Fig. A.8 (b)). Notably, some correctly classified pneumonia cases display low or diffuse attribution, indicating potential model uncertainty. This may reflect class imbalance or a reliance on the presence of identifiable structures in normal scans, rather than direct detection of pathology.

Q4.3 Consistency. In normal cases, most attribution maps appear similar, whereas pneumonia cases exhibit higher variability. In pneumonia cases the focus is predominantly on position markers ("R") and tubes or other parts of medical devices, the model is may relying on shortcuts instead of directly evaluating lung tissue.

Q4.2 Method Comparison. Grad-CAM provides smooth, region-based visualizations that are intuitive for clinical interpretation by highlighting where the model is focusing at a higher, feature map level, but it may miss subtle differences between similar-looking features. In contrast, Integrated Gradients offers pixel-level attributions that capture fine details, distinguishing between similar features based on pixel values.

2.5 Q5: Data Randomization Test

Q5.1 Retrained CNN. The CNN trained on data with randomized labels exhibited poor classification performance, similar to chance as expected.

Q5.2 & Q5.3 Lack of Explainability. Both Integrated Gradients and Grad-CAM produced attribution maps lacking meaningful or clinically relevant patterns. These maps appeared diffuse and uncertain, and in some cases, focused on irrelevant elements such as the "R" marker indicating the right side of the X-ray (cf. Appendix Fig. A.10). These findings support the view that saliency methods can reflect the trustworthiness of our previous underlying model, as the randomized model fails to learn valid associations, the attribution methods similarly fail to highlight diagnostically informative regions.

3 Part 3: General Questions

Q1.1 – Consistency across interpretability methods. Across the **three** lenses, we applied the global LASSO coefficients, SHAP explanations for the best MLP, and NAM per-feature logits; the same small clique of predictors continues to appear. *ST-slope* dummies (especially `Flat`), *Chest-pain type* dummies (`ASY`, `ATA`), *Sex_M*, *Oldpeak*, *Exercise-angina*, and (with opposite sign) *Cholesterol* dominate every ranking or importance plot. Their relative order shifts slightly, SHAP sometimes elevates an extreme *Oldpeak* or *MaxHR* value for a specific patient; however, no method highlights a variable that the others deem irrelevant, nor does any "top-tier" feature vanish elsewhere. We therefore judge the explanations to be **highly consistent**: different model classes and attribution techniques converge on the same physiological story, bolstering confidence that the patterns are signal rather than method-specific artefacts.

Q1.2 – Consistency across interpretability methods. Interpretability methods like Grad-CAM and Integrated Gradients show consistent results when analyzing normal cases due to the clear and homogeneous features of healthy lungs. However, in pneumonia cases, the consistency between methods decreases because the model's decision-making becomes more nuanced, requiring both global context and fine-grained details to capture the heterogeneous and complex patterns characteristic of the disease.

Q2.1 – How we would win a clinician's trust. We build our case on the *Neural Additive Model* (NAM), whose per-feature sub-nets $g_i(x_i)$ offer true "glass-box" visibility.
(a) Align with domain knowledge. The learned g_i curves rise for a *Flat* ST-slope, exercise-induced angina, high *Oldpeak* and male sex, and drop for an *Up* ST-slope or *ATA* chest pain, mirroring standard cardiology risk factors. Face validity is the first lever for trust.

(b) Provide patient-level accountability. For every new case the model outputs the signed contributions $\{g_i(x_i)\}$, e.g. "ST-slope contributed the most to the positive prediction, while cholesterol pulled the risk down slightly." Because each term refers to a single chart ariable, the doctor can verifyit immediately, meeting the "right to an explanation" highlighted in the lecture (slide 34).
(c) Corroborate with an independent explainer. Kernel-SHAP applied to the best MLP ranks the same top features, showing that the explanation is not an artefact of the NAM architecture.
(d) Document reliability. A one-page validation card reports 5-fold CV balanced-accuracy 0.87 ± 0.05, Brier score 0.14 and a calibration curve within ± 5 %. Performance transparency complements interpretability.
(e) Keep the workflow familiar. Translating each g_i curve into an odds-ratio lookup table lets the clinician read the model like existing risk scores, avoiding the "interpretability tax" discussed in class.

Q2.2 – How we would win a clinician's trust. To earn the trust of clinicians,cians, we need to clearly show that our attribution maps light up the right anatomical that matches their medical expertise. We should show transparent performance metrics and acknowledge limitations. Give clinicians an interactive way to explore and compare different cases and interpretability methods to prove how consistent the system is. Finally, by positioning the AI as a supportive partner that helps them, rather than replacing them, we can demonstrate real value without undermining their expertise.

Q3.1 – Do the discovered importances make clinical sense? Yes. The **same top-ranked variables** across Lasso, SHAP–MLP and NAM mirror well-established cardiology knowledge:

- **ST-slope (Flat ↑, Up ↓).** A flat recovery slope is textbook evidence of ischaemia, whereas an up-sloping trace is usually benign.
- **Chest-pain type (ASY ↑, ATA ↓).** Asymptomatic (silent) angina is a red flag; typical-angina (ATA) often reflects exertional pain without infarction risk, hence the protective weight.
- **Sex_M (↑).** Male sex is a non-modifiable risk factor with consistently higher incidence of coronary disease.
- **Oldpeak (↑).** Larger exercise-induced ST depression indicates more severe myocardial oxygen deficit.
- **Exercise-angina (↑).** Pain triggered by exercise suggests flow-limiting stenosis.
- **Cholesterol (↓).** The weak negative weight is counter-intuitive but plausible after multivariable adjustment: many high-cholesterol patients are on statins, while those with very low readings can be acutely ill. Its small magnitude (relative to ST-slope or chest pain) signals low explanatory power rather than a model failure.

Q3.2 – Do the discovered importances make clinical sense? Yes. Even without medical training, the differences between normal and pneumonia X-rays are visually apparent, and our interpretability methods highlight precisely those regions—showing clear clinical relevance. In pneumonia cases, the model focuses on affected lung tissue, whereas in normal scans it emphasizes skeletal boundaries and borders, likely because there's no abnormal lung tissue to highlight.

Q4.1 – Which model would we put in production, and why? We would **deploy the leak-proof *Logistic-Lasso* pipeline** (14 $\rightarrow$ 12 active features).

(i) **Competitive and stable performance.** CV balanced-accuracy 0.87 matches the more complex MLP and exceeds the NAM's 0.81, with narrower error bars and no training instability.

(ii) **Full intrinsic interpretability.** Each coefficient is a direct log-odds multiplier; no post-hoc tooling or surrogate plots are needed to audit decisions, critical for regulatory approval and the "right-to-explanation" requirement.

(iii) **Parsimony.** Twelve routinely collected variables keep data entry cost low and reduce missing-value headaches; the model fits in a single SQL statement or an EHR rules engine.

(iv) **Ease of maintenance.** Updating coefficients with new data is a one-line re-fit; recalibration (e.g. via Platt scaling) is textbook and fast.

(v) **Regulatory clinical acceptance.** Logistic models are well understood by hospital ethics boards, easier to validate against prospective cohorts, and can be expressed as a lookup nomogram, minimising the "interpretability tax" highlighted in the lecture.

(vi) **Lightweight deployment.** No GPU/TPU, no third-party libraries, millisecond latency: suitable for bedside tablets or embedded devices.

The NAM is a close runner-up for its richer per-feature curves, but its lower accuracy and higher computational footprint tip the balance toward the simpler, trust-ready Logistic-Lasso.

Q4.2 – Which model would we put in production, and why? We would put the CNN variant in production that achieves the best balance between predictive performance, robustness, and interpretability. Additionally, we propose using Grad-CAM as the primary interpretability tool, because it is computational efficiency, while reserving Integrated Gradients for cases demanding detailed, pixel-level attribution, in an interactive framework that permits direct comparison of both methods.

References

[1] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[2] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, and G. Hinton, "Neural additive models: Interpretable machine learning with neural nets," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

A Appendix / Supplementary Material

Exploratory-data details for Part 1

Table A.1: **Numeric feature characteristics and planned fixes.**

Feature	Empirical range / shape	Bad vals	Handling strategy
Age	$29-77$, near–Gaussian	—	z-scale only
RestingBP	$80-200$ mmHg; spike 0	50 zeros	$0 \rightarrow$ NaN, KNN(k=5)
Cholesterol	$70-603$ mg/dl; heavy tail	170 zeros	$0 \rightarrow$ NaN, KNN; keep tail
MaxHR	$60-202$ bpm; unimodal	—	z-scale only
Oldpeak	$0-6.2$; mass 0	9 negatives	$<0 \rightarrow$ NaN, floor 0

Table A.2: **Categorical level proportions (train–val split).**

Column	Level distribution (% of rows)
Sex	M 78, F 22
ChestPainType	ASY 52, NAP 23, ATA 21, TA 4
RestingECG	Normal 60, LVH 20, ST 20
ExerciseAngina	N 60, Y 40
ST_Slope	Flat 50, Up 44, Down 6

Pre-processing blueprint (scikit-learn object) for Part 1

```
ColumnTransformer({
  num: Pipeline({StandardScaler → KNNAndScaledBounds(k=5) → StandardScaler}),
  cat: Pipeline({SimpleImputer(most_frequent) → OneHotEncoder(drop='first', handle_unknown='ignore')})
})
```

Note. The object above is stored un-fitted as `preprocessor.pkl`. Every downstream model clones and fits it inside its own cross-validation fold, thereby preventing data leakage.

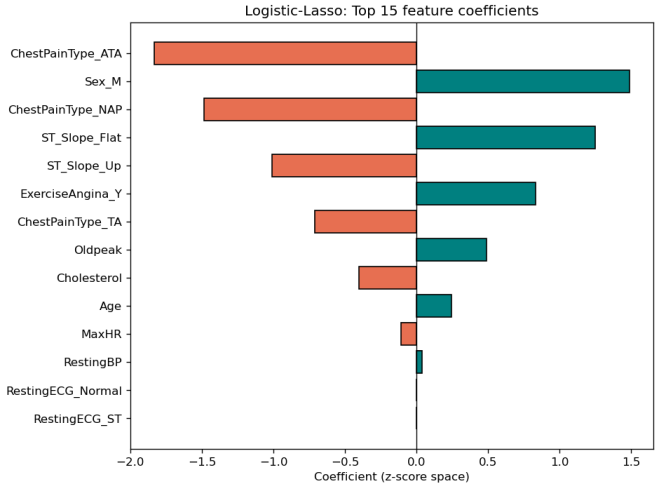
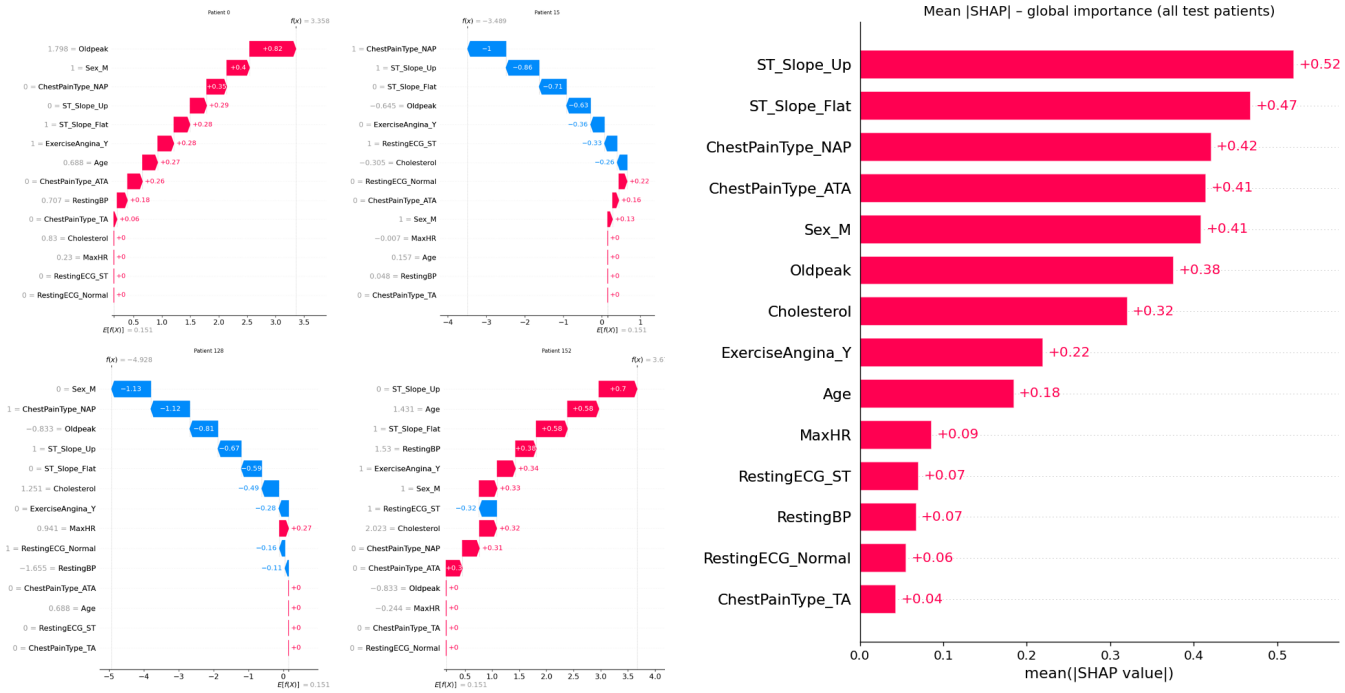


Figure A.2: **Logistic-Lasso coefficients.** Positive values (teal) increase, negative (coral) decrease the log-odds of *HeartDisease*=1. Only the 15 strongest effects are shown; two ECG dummies are exactly zero.



((a)) Local waterfalls ((b)) Global importance

Figure A.3: **SHAP explanations for the best-MLP.** (a) Four representative patients; (b) population-level importance.

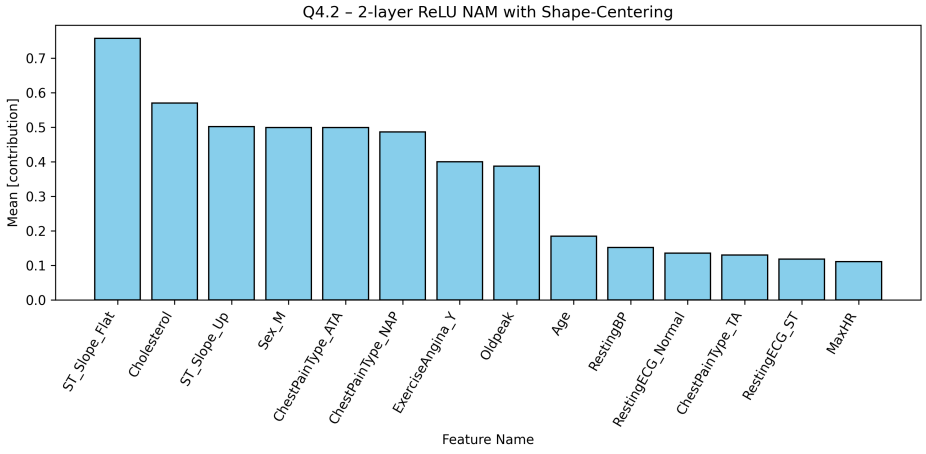


Figure A.4: **NAM global importance.** Mean $|g_i|$ across the test set.

CNN Confusion Matrix Part 2

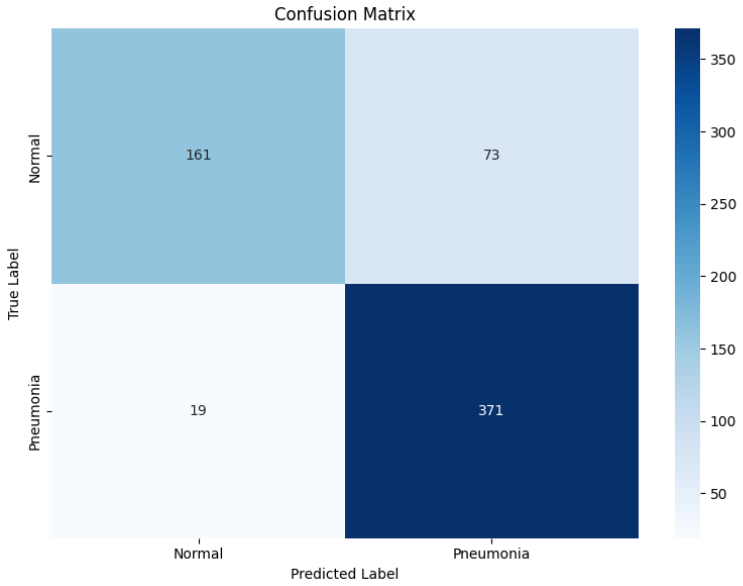


Figure A.5: (Confusion matrix showing the performance of the CNN in classifying chest X-rays as either normal or pneumonia cases. The model correctly classified 161 normal and 371 pneumonia cases, while misclassifying 73 normal cases as pneumonia and 19 pneumonia cases as normal. This indicates strong performance in pneumonia detection with some false positives.

Integrated Gradient Attribution Maps for Part 2

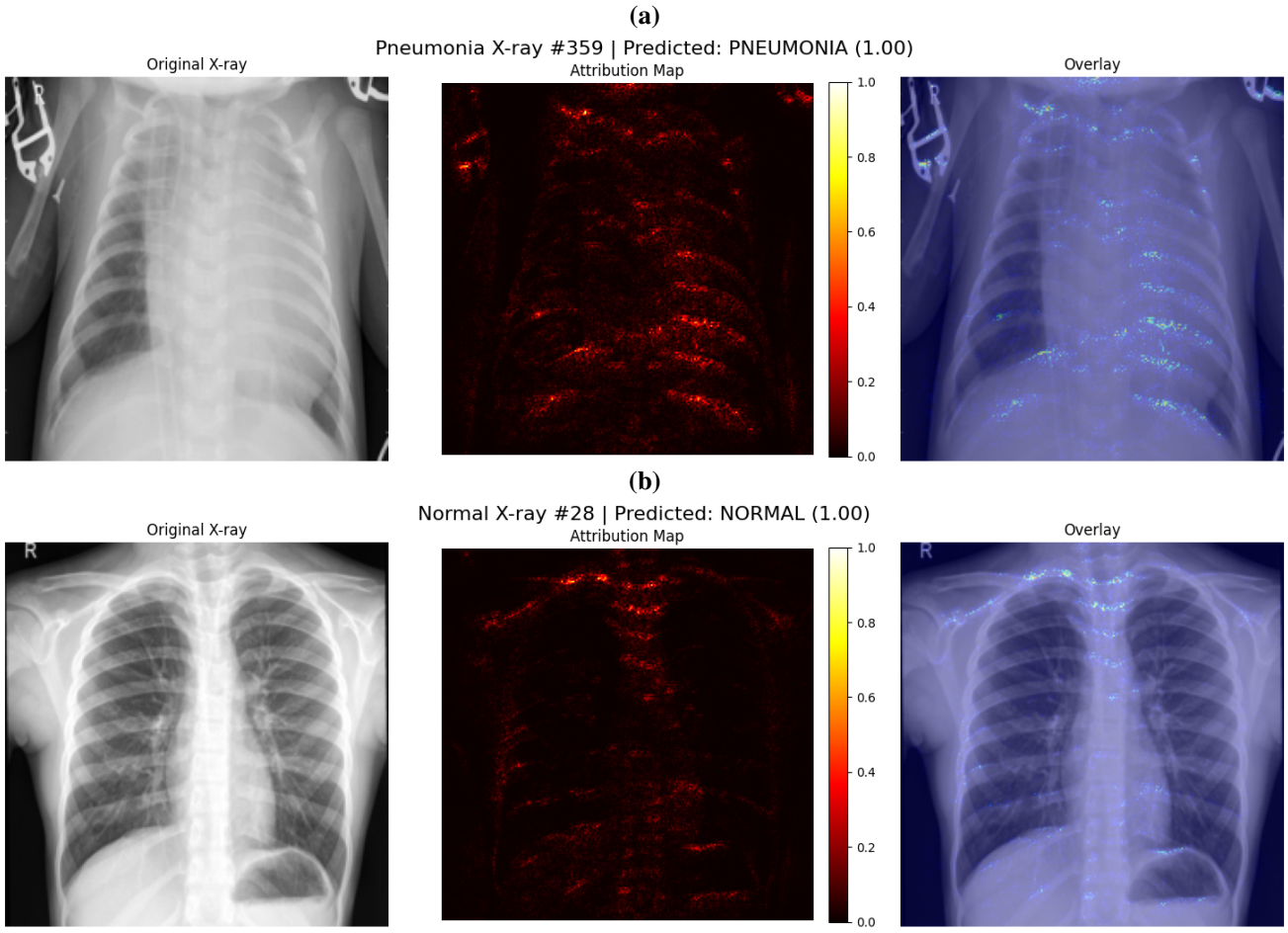


Figure A.6: (a) Integrated Gradient attribution map for a pneumonia case. (b) Integrated Gradient attribution map for a normal case.

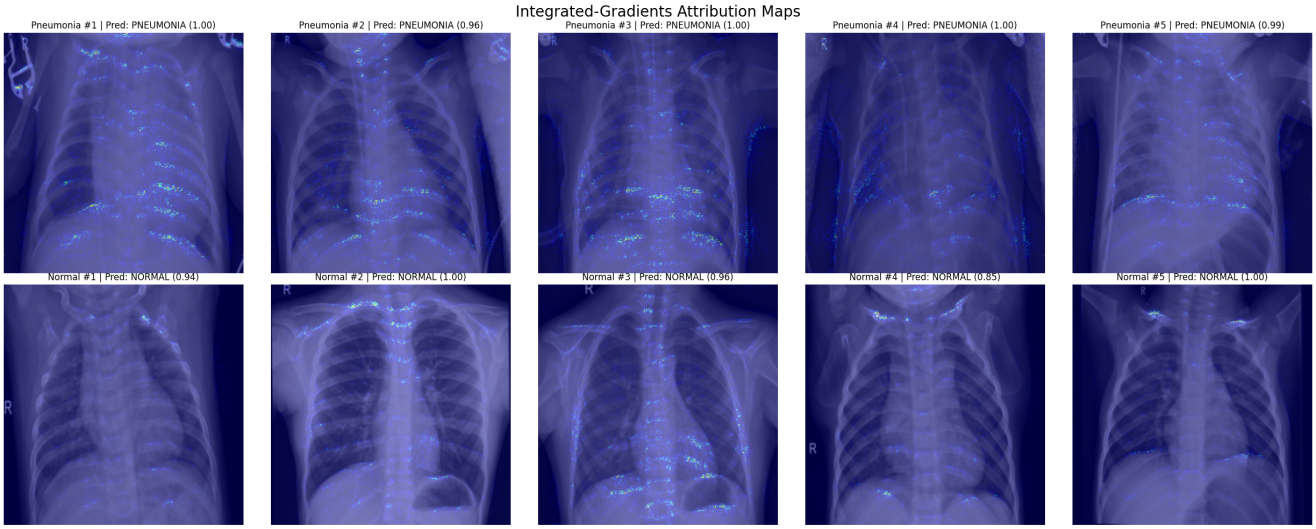


Figure A.7: Overlap between X-ray image and Integrated Gradients attribution map for five normal and five pneumonia samples.

Grad-CAM Attribution Maps for Part 2

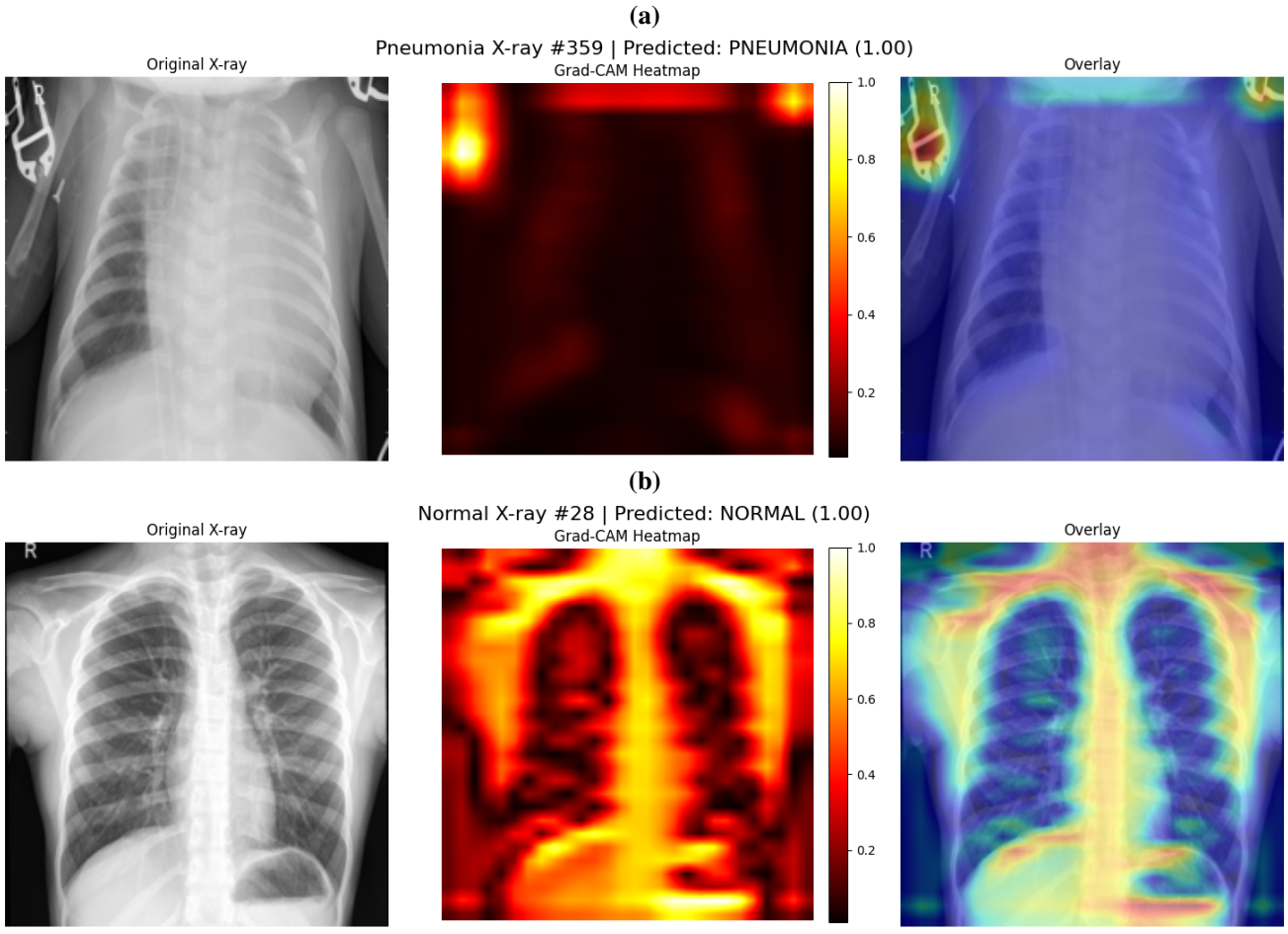


Figure A.8: (a) Grad-CAM attribution map for a pneumonia case. (b) Grad-CAM attribution map for a normal case.

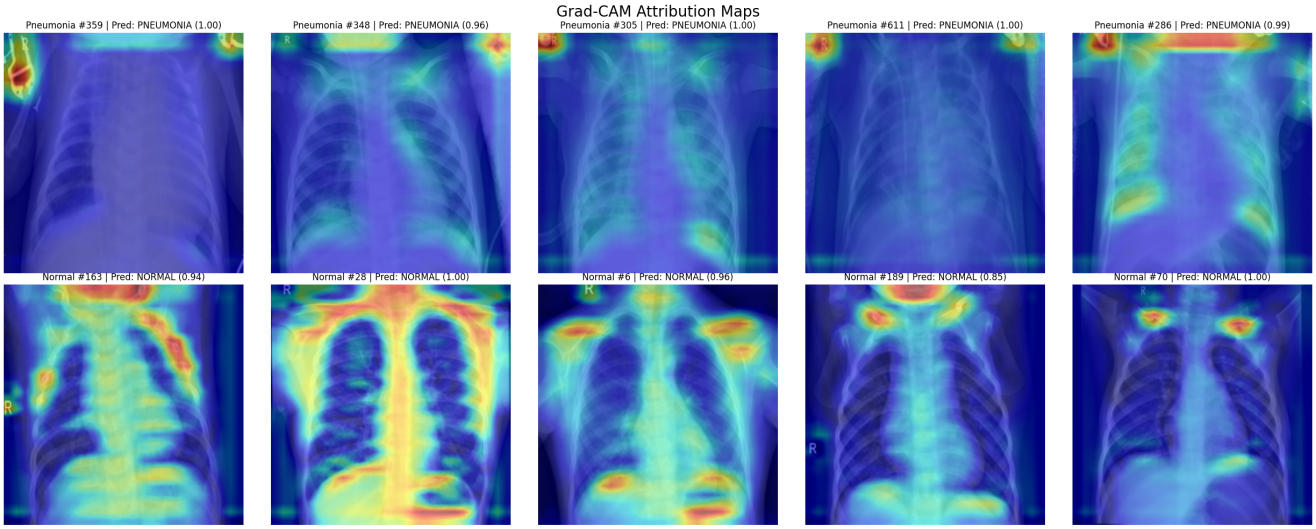


Figure A.9: Overlap between X-ray image and Grad-CAM attribution map for five normal and five pneumonia samples.

Attribution Maps Randomized Model for Part 2

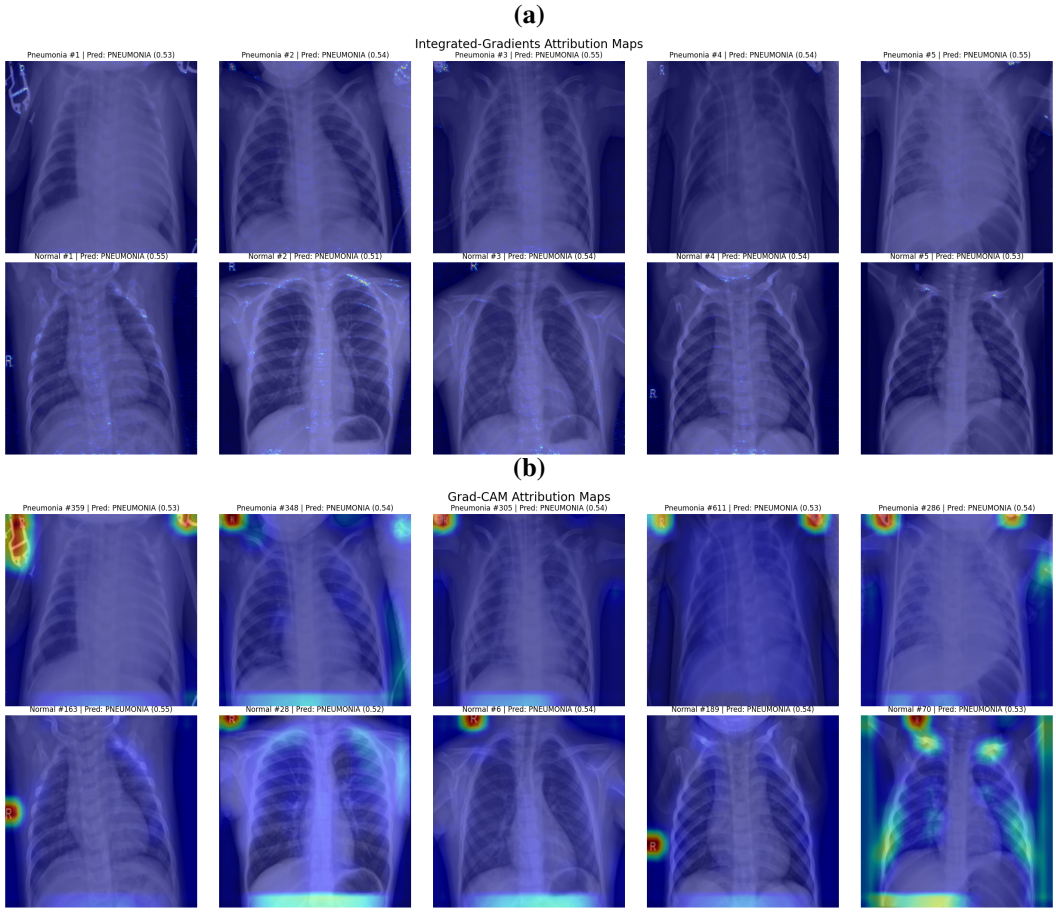


Figure A.10: (a) Integrated Gradients attribution maps from randomized model. (b) Grad-CAM attribution maps from randomized model. Neither methods generated attribution maps that revealed clinically meaningful patterns based on the randomized model.