PULSE-LAB: A Multimodal Hybrid State-Space Model for Forecasting the Presence of Thoracic Pathologies from ECG Time Series and Laboratory Data

Anonymous Author(s)

Abstract

In acute care, a gap often exists between admission and confirmatory imaging. We address this gap by forecasting chest X-ray (CXR) findings hours in advance using admission-time biosignals alone. We present PULSE–LAB, a hybrid model that combines a Mamba state-space encoder for long-sequence ECG with an MLP over 50 routine laboratory tests. Using the public Symile–MIMIC dataset, we predict 14 thoracic findings on the first post-admission CXR and establish, to our knowledge, the first benchmark for this task. The model achieves a macro-AUROC of 0.62 and macro-recall of 0.63, with strongest discrimination for acute conditions (e.g., Fracture, AUROC 0.81) and high recall for developing pulmonary processes (e.g., Pneumonia, Consolidation). Interpretability analyses reveal clinically coherent patterns: coagulation and injury markers drive fracture risk, while renal and oncotic markers relate to edema. Both the ordered lab panel and informative ECG leads contribute. These results indicate that admission-time signals capture early physiological trajectories that later manifest on imaging, suggesting utility for earlier risk stratification, imaging prioritization, and proactive care.

1 Introduction

Accurate, timely diagnosis of thoracic pathologies in emergency care is complex, and clinical decisions are often made before confirmatory chest X-rays (CXRs) are obtained [1; 2; 3; 4]. Delays arise from logistics, resource constraints, and patient instability [5; 6; 7; 8].

We investigate whether admission-time biosignals can forecast future radiographic findings, reframing diagnosis from retrospective interpretation to prospective prediction. Specifically, we forecast the presence of specific CXR findings several hours ahead using only a 10-second 12-lead ECG and routine laboratory tests available at admission.

Our main contributions are: (1) We define a new forecasting task: predicting 14 CXR findings on the first post-admission image from admission-time ECG and labs. (2) We propose PULSE–LAB, a hybrid architecture with a Mamba ECG encoder and an MLP for tabular labs. (3) We establish the first benchmark on Symile–MIMIC, achieving macro-AUROC 0.62 with strong performance for acute conditions (e.g., Fracture 0.81), highlighting potential for triage and imaging prioritization.

2 Related Work

Early work on chest X-ray (CXR) interpretation applied convolutional neural networks (CNNs) to retrospective image classification, establishing the feasibility of automated CXR analysis [9]. These approaches remain retrospective, requiring a CXR image prior to analysis and thus cannot predict future developments.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

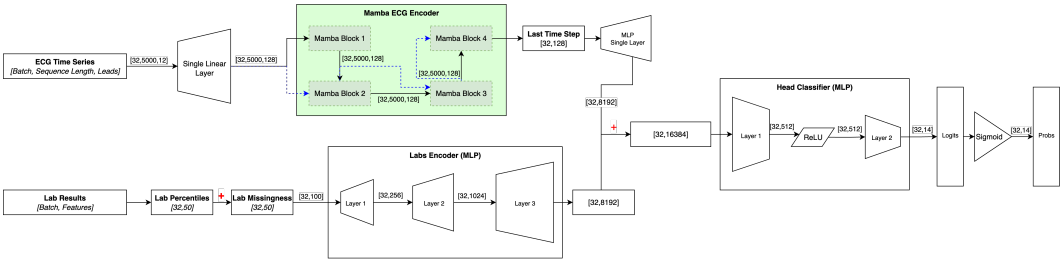

Figure 1: The PULSE–LAB: a Mamba encoder processes the ECG; an MLP encodes lab values and missingness. Embeddings are concatenated and fed to a classifier to forecast 14 CXR findings.

Another research direction has focused on forecasting outcomes using non-imaging biosignals, such as data from electronic health records (EHRs) [10; 11]. These findings suggest that biosignals contain predictive information about thoracic conditions. However, existing models either focus on general outcomes rather than specific image-based pathologies or remain retrospective in nature. Our work is directly inspired by the Symile-MIMIC dataset [12], which demonstrated that a shared signal exists across modalities. Our study refocuses on the more complex task of diagnostic forecasting, a key clinical application at the intersection of time series analysis and health. Rather than matching patients' laboratories and X-rays, we predict 14 specific radiographic findings based on admission-time biosignals.

3 Methodology

3.1 Dataset and Task Formulation

This study uses the public **Symile–MIMIC** dataset [12], a multimodal collection of 11,622 unique hospital admissions from 9,573 patients, with pre-defined data splits (10,000 train, 750 validation, 464 test) to prevent data leakage. The dataset integrates components from the PhysioNet ecosystem: *MIMIC-IV* (demographics and laboratory results), *MIMIC-IV-ECG* (electrocardiograms), and *MIMIC-CXR-JPG* (radiographs with CheXpert labels).

Each admission provides two modalities recorded within the first 24 hours:

- **ECG ($\mathbf{x_{ECG}}$):** A 10-second, 12-lead signal sampled at 500 Hz, represented as a 12×5000 matrix normalized to the range $[-1, 1]$.
- **Labs ($\mathbf{x_{labs}}$):** A panel of 50 blood tests, forming a sparse and irregularly-sampled dataset, a key data-specific challenge. Values are standardized as percentiles of the training set's empirical CDF, and each is paired with a binary missingness indicator, forming a 100-dimensional vector.

The predictive target is a 14-dimensional binary vector $\mathbf{y}$ derived from the CheXpert labeler annotations of the first post-admission CXR. Following the dataset's protocol, uncertain labels (-1) are mapped to negative (0). The forecasting task is to learn a function $f : (\mathbf{x}_{\text{ECG}}, \mathbf{x}_{\text{labs}}) \mapsto \mathbf{z} \in \mathbb{R}^{14}$, where $\mathbf{z}$ is a logit vector. The final probabilities are obtained via $\hat{\mathbf{y}} = \sigma(\mathbf{z})$.

3.2 Model Architecture: PULSE–LAB

We propose PULSE–LAB, a dual-encoder architecture designed to process the distinct data types of ECGs (long, dense time series) and labs (sparse, tabular features), as shown in Figure 1.

ECG Encoder (Mamba SSM) To capture long-range dependencies in the 5000-timestep ECG sequences with linear complexity, we use a Mamba state-space model (SSM) [13], a novel architecture well-suited for such signals. The input is first projected from 12 leads to a hidden dimension $d_{\text{model}} = 128$. A stack of four Mamba blocks processes this sequence, each configured with a state dimension $d_{\text{state}} = 16$, a convolutional kernel of size 4, and an expansion factor of 2. We use the final timestep representation of the sequence output, which is then projected to a shared embedding dimension of $d = 8192$ and normalized with LayerNorm to produce $\mathbf{E}_{\text{ECG}}$. This embedding dimension was adopted from the original Symile–MIMIC study for comparability [12].

Table 1: Forecasting performance of PULSE-LAB. Results are macro-averaged across 14 findings.

Pathology	AUROC	Recall
Fracture	0.81	0.70
Lung Lesion	0.70	0.58
Support Devices	0.67	0.63
No Finding	0.65	0.42
Edema	0.64	0.65
Atelectasis	0.63	0.65
Consolidation	0.62	0.79
Pleural Effusion	0.61	0.68
Pleural Other	0.60	0.75
Lung Opacity	0.59	0.73
Pneumonia	0.59	0.74
Cardiomegaly	0.57	0.72
Pneumothorax	0.50	0.35
Enlarged Cardiomediastinum	0.48	0.43
Average (Macro)	0.62	0.63

Laboratory Encoder (MLP) A three-layer multilayer perceptron (MLP) encodes the 100-dimensional laboratory vector. The network follows a $100 \to 256 \to 1024 \to 8192$ transformation with GELU activations and a final layer normalization (LayerNorm) layer. This design allows the model to learn non-linear relationships between laboratory values while also learning from the clinical decisions about which tests were performed, as captured by the missingness indicators.

Fusion and Classification The modality embeddings are concatenated to form a fused vector $\mathbf{h} = [\mathbf{E}_{\text{ECG}} \,\|\, \mathbf{E}_{\text{LAB}}] \in \mathbb{R}^{2d}$. This vector is passed through a bottleneck classifier MLP ($16384 \to 512 \to 14$) with a ReLU activation and Dropout (0.2) for regularization. Simple concatenation was chosen after experiments with cross-attention fusion proved less stable and yielded no performance gains.

3.3 Training, Evaluation, and Interpretability

The model is implemented in PyTorch Lightning and trained for 50 epochs on a single NVIDIA A100 GPU with a batch size of 32. We use the AdamW optimizer with a learning rate of 1×10^{-4}, betas of $(0.9, 0.999)$, and weight decay of 0.01. The loss function is a weighted binary cross-entropy with logits:

$$\mathcal{L}(\mathbf{z}, \mathbf{y}) = -\sum_{c=1}^{14} \Big[w_c \, y_c \log \sigma(z_c) + (1 - y_c) \log(1 - \sigma(z_c)) \Big],$$

where class-specific weights $w_c = N_c^{\text{neg}} / N_c^{\text{pos}}$ are computed from the training set label frequencies to compensate for class imbalance.

Performance is evaluated using macro-averaged AUROC and macro-averaged Recall. Macro-averaging ensures that performance on rare but clinically critical findings is not outweighed by common ones. The model checkpoint with the highest validation macro-AUROC is used for final testing. As a component of developing trustworthy AI, we explain model predictions using SHAP (SHapley Additive exPlanations) to attribute feature importance for the laboratory panel and assess the contribution of each of the 12 ECG leads.

4 Results

Forecasting Performance PULSE-LAB achieved a macro-averaged AUROC of 0.62 and a macro-averaged recall of 0.63. This result confirms the feasibility of forecasting future radiographic findings from sparse, admission-time biosignals. Performance varies across pathologies, as shown in Table 1 and Figure 2. The model performs best on acute conditions with systemic physiological footprints, such as Fracture (AUROC 0.81), Lung Lesion (0.70), and Support Devices (0.67). For developing pulmonary pathologies, the model achieves moderate AUROCs but high recall, suggesting its utility for screening. Our model also identifies patients who will later show Consolidation (Recall 0.79), Pneumonia (Recall 0.74), Pleural Effusion (Recall 0.68), and Edema (Recall 0.65) with high sensitivity. Conditions with weak precursor signals, such as the chronic state of Cardiomegaly (AUROC 0.57) or the abrupt onset of Pneumothorax (AUROC 0.50), remain difficult to forecast.

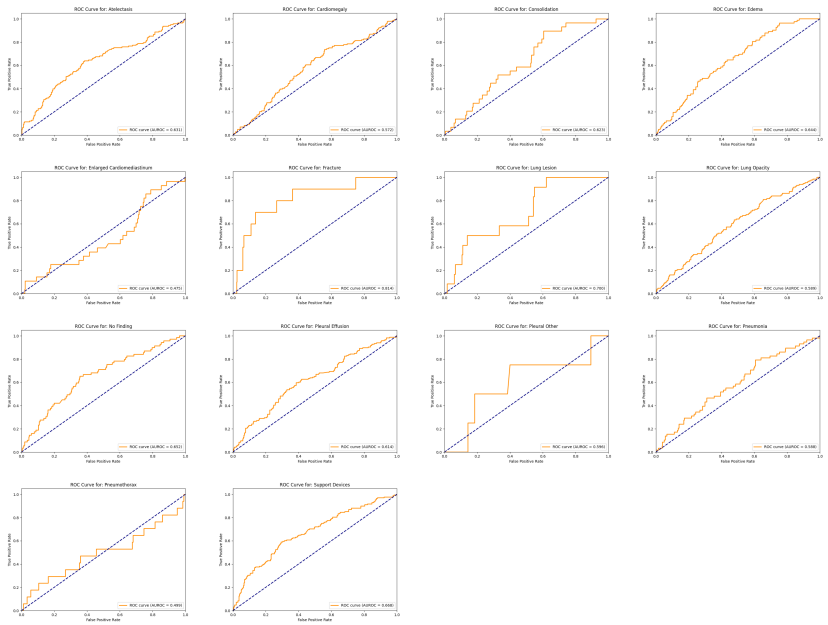

Figure 2: ROC curves for all 14 findings. Strong discrimination for acute processes (e.g., *Fracture*); near-random for abrupt/weak-precursor cases (e.g., *Pneumothorax*).

Model Interpretability We used SHAP feature attributions to understand the model's predictions (see Appendix A). The analysis reveals two key mechanisms. First, the MLP encoder learns clinically coherent, pathology-specific laboratory profiles. For example, fracture risk is driven by markers of injury and coagulation (e.g., elevated 'Creatine Kinase', 'INR', 'PTT') [14], while edema risk is associated with markers of renal dysfunction and fluid overload (e.g., high 'Urea Nitrogen', low 'Albumin') [15; 16]. The model also learns from missingness flags, using the pattern of which tests were ordered as a proxy for clinical suspicion. Second, the Mamba encoder learns a generalized signature of patient acuity from the ECG. Across all pathologies, leads fundamental for assessing heart rate, rhythm, and electrical axis (I, II, aVF, V2) consistently show the highest importance. This suggests the ECG provides a baseline of cardiorespiratory stress, which contextualizes the specific signals from the lab panel.

5 Discussion

Forecasting CXR findings from admission-time biosignals is tractable yet challenging. A macro-AUROC of 0.62 confirms that early signals contain trajectories that culminate in radiographic abnormalities. The pattern is clinically coherent: strong performance for acute, systemic processes (e.g., Fracture) and high recall for developing pulmonary disease (e.g., Pneumonia), supporting use as a screening and prioritization aid.

Interpretability links model behavior to pathophysiology, utilizing coagulation/injury markers for fracture, renal/oncotic markers for edema, and ECG leads that capture a broad acuity signal complementing targeted laboratory information. Such mechanisms are consistent with clinical reasoning and can inform earlier risk stratification and the allocation of imaging.

Limitations include reliance on a single dataset and conservative handling of uncertain labels (mapped to negative). Future work should test multi-center generalization and alternative label strategies.

In conclusion, PULSE–LAB shifts from retrospective diagnosis to prospective forecasting using signals already available at admission, enabling earlier, more proactive care.

References

[1] Thi Dan Linh Nguyen-Kim. Imaging of chest trauma. In *Diseases of the Chest, Heart and Vascular System 2025-2028: Diagnostic Imaging*, pages 109–134. Springer, 2025.

[2] Sirote Wongwaisayawan, Ruedeekorn Suwannanon, Sorravit Sawatmongkorngul, and Rathachai Kaewlai. Emergency thoracic us: the essentials. *Radiographics*, 36(3):640–659, 2016.

[3] Yaser Mohammed Al-Worafi. Radiology and medical imaging research in developing countries. In *Handbook of Medical and Health Sciences in Developing Countries: Education, Practice, and Research*, pages 1–29. Springer, 2024.

[4] Fernanda Bellolio, Michael Gottlieb, Richard Body, Martin P Than, and Erik P Hess. Evaluating patients with chest pain in the emergency department. *bmj*, 388, 2025.

[5] S Cournane, R Conway, D Creagh, DG Byrne, N Sheehy, and B Silke. Radiology imaging delays as independent predictors of length of hospital stay for emergency medical admissions. *Clinical radiology*, 71(9):912–918, 2016.

[6] Andrew Worster, Christopher MB Fernandes, Cheryl Malcolmson, Kevin Eva, and Diane Simpson. Identification of root causes for emergency diagnostic imaging delays at three canadian hospitals. *Journal of Emergency Nursing*, 32(4):276–280, 2006.

[7] Amir A Khaliq, Eugene Nsiah, Nadia H Bilal, Danny R Hughes, and Richard Duszak Jr. The scope and distribution of imaging services at critical access hospitals. *Journal of the American College of Radiology*, 11(9):857–862, 2014.

[8] Lora K Ott, Marilyn Hravnak, Sunday Clark, and Nikhil B Amesur. Patients' instability, emergency response, and outcomes in the radiology department. *American Journal of Critical Care*, 20(6):461–469, 2011.

[9] Uswatun Hasanah, Cries Avian, J. T. Darmawan, Nabil Bachroin, Muhamad Faisal, Setya Widyawan Prakosa, Jenq-Shiou Leu, and Chia-Ti Tsai. Chexnet and feature pyramid network: a fusion deep learning architecture for multilabel chest x-ray clinical diagnoses classification. *The international journal of cardiovascular imaging*, 2023.

[10] Yangyi Song, Hao Ren, Fengshi Jing, Chaocheng He, Yewei Xie, and Jiandong Zhou. Enhancing pneumonia diagnosis using a ensemble transformer-based attention network for chest x-ray image analysis. In *2024 2nd International Conference on Big Data and Privacy Computing (BDPC)*, pages 154–159, 2024.

[11] Md. Abu Sufian, Wahiba Hamzi, Tazkera Sharifi, Sadia Zaman, Lujain Alsadder, Esther Lee, Amir Hakim, and Boumediene Hamzi. Ai-driven thoracic x-ray diagnostics: Transformative transfer learning for clinical validation in pulmonary radiography. *Journal of Personalized Medicine*, 14, 2024.

[12] Adriel Saporta, Aahlad Manas Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. *Advances in Neural Information Processing Systems*, 37:56919–56957, 2024.

[13] Tri Dao and et al. Mamba: Linear-time sequence modeling with selective state spaces. In *International Conference on Learning Representations (ICLR)*, 2023.

[14] Ke Wen, Zhexuan Lin, Haizhu Tan, and Ming Han. Correlations between coagulation abnormalities and inflammatory markers in trauma-induced coagulopathy. *Frontiers in Physiology*, 15:1474707, 2024.

[15] Rolando Claure-Del Granado and Ravindra L Mehta. Fluid overload in the icu: evaluation and management. *BMC nephrology*, 17(1):109, 2016.

[16] Angelo Gatta, Alberto Verardo, and Massimo Bolognesi. Hypoalbuminemia. *Internal and emergency medicine*, 7(Suppl 3):193–199, 2012.

A Appendix: Model Interpretability Analysis

This appendix provides detailed SHAP attribution analyses that were summarized in the main paper.

A.1 Laboratory Feature Attributions

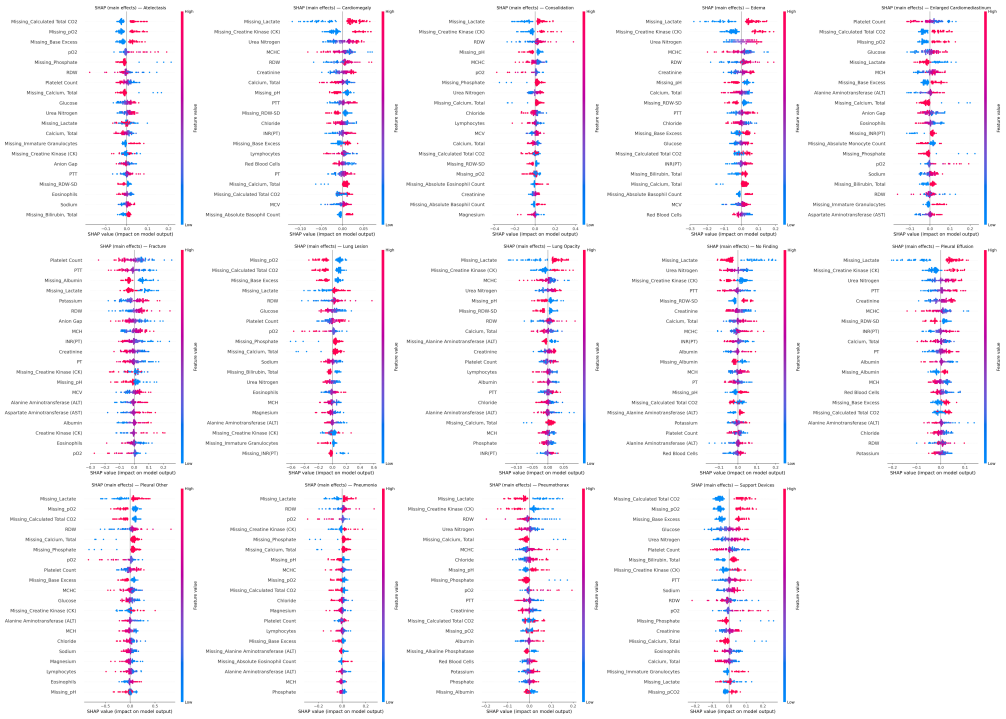

Figure 3: Laboratory SHAP beeswarm plots across CXR findings. Each dot represents a patient: the horizontal position indicates the SHAP effect on the logit, and the color represents the feature value (*red = high*, *blue = low*).

Interpreting SHAP Beeswarm Plots (Figure 3) The beeswarm plots summarize per-patient SHAP values for each laboratory test. Each dot represents a single patient's feature attribution for a specific disease or condition. The horizontal position indicates the SHAP value's impact on the model's output logit (positive values push the prediction towards 1, negative values towards 0). The color of the dot indicates the feature's value, from low (blue) to high (red). This visualization reveals not only which features are important but also how their values relate to risk. For example, for Edema, high 'Urea Nitrogen' values (red dots) cluster on the positive SHAP side, indicating that azotemia increases the predicted risk. Similarly, low 'Albumin' values (blue dots) also push the prediction in a positive direction, reflecting the clinical link between hypoalbuminemia and fluid overload. The model also learns from missingness flags; for Atelectasis, the presence of a pO2 measurement (blue dots for Missing_pO2) is associated with a higher risk, suggesting that the clinical decision to order an arterial blood gas test itself serves as a predictive signal.

Interpreting Top-20 Feature Importance (Figure 4) This figure aggregates the mean absolute SHAP values for each laboratory feature across all test patients, showing the top 20 most influential labs for each pathology. It highlights that the model learns distinct, task-specific laboratory profiles rather than relying on a single dominant test. For Fracture, coagulation, and injury markers like 'INR(PT)', 'PTT', and 'Creatine Kinase' are most influential. For Consolidation and Pneumonia, markers of hypoxemia and acid-base imbalance ('pO2', 'pH') are key. This demonstrates that the model discovers clinically plausible, multifaceted biomarker signatures for each condition.

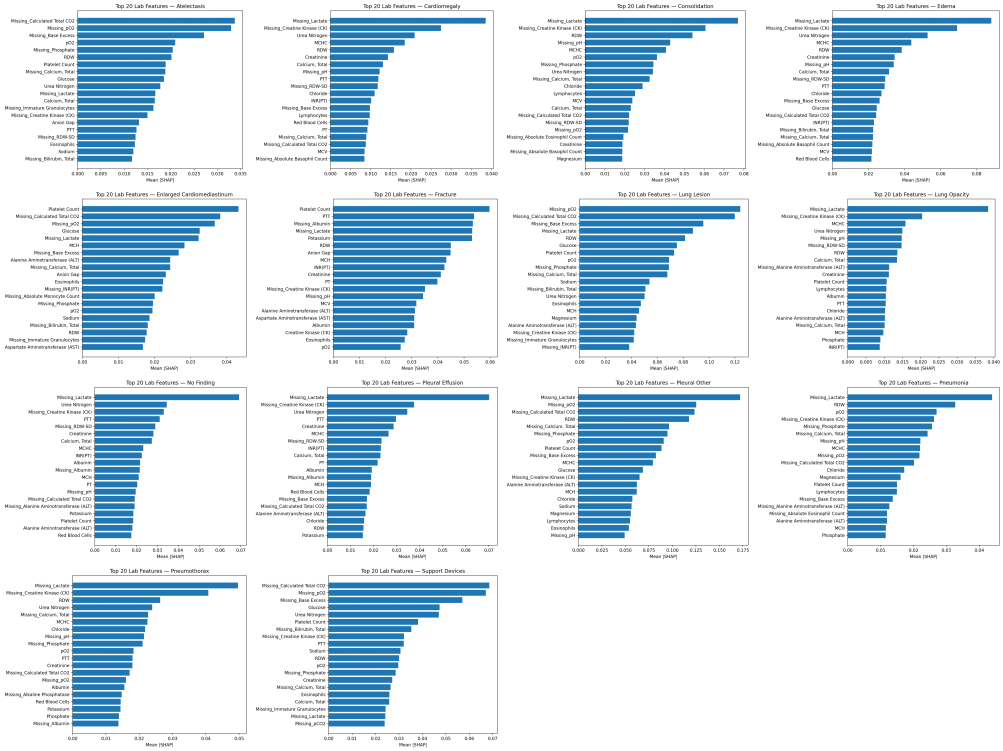

Figure 4: Top-20 laboratory features per pathology, ranked by mean absolute SHAP value.

A.2 ECG Lead-Level Attributions (Figure 5)

Figure 5 shows the mean absolute SHAP values for each of the 12 ECG leads across all pathologies. The analysis reveals a highly stable importance profile, where leads I, II, aVF, and V2 consistently show the highest contribution, regardless of the target pathology.

A plausible interpretation is that the ECG encoder learns a generalized signature of overall patient acuity and cardiorespiratory stress, rather than distinct features for each pathology. Diverse conditions, from trauma (Fracture) to infection (Pneumonia), can trigger common systemic responses such as tachycardia or shifts in the heart's electrical axis. The identified leads are important for assessing rate, rhythm, and the frontal plane axis, making them ideal for capturing such a non-specific "acuity" signal. The magnitude of the SHAP values varies by pathology, likely reflecting the degree to which each condition induces systemic stress. This suggests the ECG provides a contextual baseline of patient severity, which complements the more diagnostic-specific information from the laboratory panel.

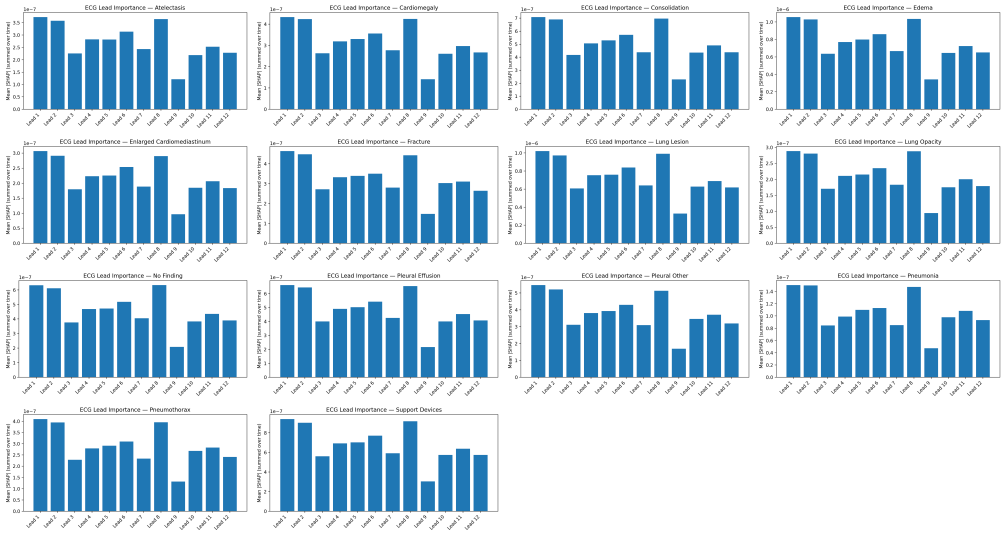

Figure 5: ECG lead-level attributions summarized by mean absolute SHAP values.