Architecture-Aware Explainability in ECG Analysis: A Case Study of Aortic Stenosis Detection with ResNet18, LSTM and ViT-MAE ECG

Samuel Chol Buol[1], **Julius Zannu**[1], **Carine Pierrette Mukamakuza**[1], **Damilare Emmanuel Olatunji**[1], **Vijayakumar Bhagavatula**[2]

[1] Carnegie Mellon University Africa
Kigali, Rwanda
[2] Carnegie Mellon University
Pittsburgh, United States of America
{sbuol, jzannu, cmukamakuza, dolatunji}@andrew.cmu.edu, kumar@ece.cmu.edu

Abstract

Aortic stenosis (AS) remains a major cardiovascular challenge, as early diagnostic markers in electrocardiogram (ECG) signals are often subtle and difficult to identify using conventional approaches. Although deep learning models have demonstrated strong performance in AS detection, their clinical adoption is limited by the insufficient interpretability of model decisions. Existing explainability studies typically focus on individual architectures, leaving open the question of whether different model designs rely on distinct ECG features. In this work, we investigate how the architecture of neural networks influences the explainability and clinical interpretability in the classification of AS based on ECG. We systematically compare three architectures, namely ResNet18, Long Short-Term Memory (LSTM), and a Vision Transformer with Masked Autoencoder (ViT-MAE),trained in the open-access Cardio-mechanical Signals database comprising 100 patients with valvular heart diseases. All models achieved strong predictive performance, with accuracies of 97.23% (ResNet18), 98.96% (LSTM), and 88.56% (ViT-MAE). To analyze model behavior, we apply both Integrated Gradients and Local Interpretable Model-agnostic Explanations (LIME). The results reveal architecture-specific attribution patterns: ResNet18 exhibits a broad attention across P-waves, QRS complexes, and ST–T segments; LSTM emphasizes temporally salient QRS related features; and ViT-MAE prioritizes repolarization associated regions, including T-waves and QT intervals. Despite these differences, all architectures consistently focus on clinically meaningful ECG regions associated with AS pathophysiology. These findings demonstrate that explainability outcomes are strongly influenced by model architecture and underscore the importance of architecture-aware interpretability strategies for building transparent, reliable and clinically trustworthy AI systems for cardiovascular diagnosis.

Code — https://github.com/choldit/architecture-aware-explainability

Datasets — https://zenodo.org/records/5279448

Introduction

Each year, nearly one-third of all global deaths (around 17.9 million in recent estimates, (WHO 2025; GBD Cardiovascular Disease Collaboration 2023)) are caused by cardiovascular disease (CVD). (Institute for Health Metrics and Evaluation 2023). Among these conditions, aortic stenosis (AS) presents a particularly difficult screening problem, as the standard 12-lead electrocardiogram (ECG) often fails to detect early disease due to subtle electrical changes (Thoenes et al. 2018). Traditional criteria exhibit low sensitivity in mild AS, meaning many patients progress undiagnosed until symptoms emerge (van Geldorp et al. 2009; Iung et al. 2003). This diagnostic gap places a growing emphasis on accessible tools. Deep learning applied to ECGs has shown promise: for example, an AI-ECG model achieved an 0.884 (95% CI, 0.880–0.887) and 0.861 (95% CI, 0.858–0.863), respectively; those using a single-lead ECG signal were 0.845 (95% CI, 0.841–0.848) and 0.821 (95% CI, 0.816–0.825), respectively over 39,000 ECG records. Equally, (Aminorroaya et al. 2023)'s model achieved AUROC of 0.829 (95 % CI: 0.800-0.855) for moderate/severe AS and 0.846 (95 % CI: 0.778-0.899) for severe AS, with sensitivity of 90.4 % and specificity of 58.7 % at the optimized threshold, yielding NPV of 99.2 % at the observed 4.5 % prevalence. These results suggest the potential of ECG-based screening for AS, yet the path to clinical integration remains incomplete.

Despite technical advances, the major barrier to clinical adoption is interpretability: deep neural networks are often opaque, limiting trust and regulatory acceptance. In cardiology, where decisions carry high patient risk, clinicians demand transparency to ensure AI outputs reflect physiological phenomena rather than artifacts. Explainable AI (XAI) methods-such as SHAP, Grad-CAM, and Integrated Gradients—are increasingly used in cardiac settings to visualise or attribute model decisions (Kwon et al. 2020). (Purwono, Wulandari, and Nisa 2025). However, these efforts typically focus on a single architecture (e.g.CNN variants) and do not investigate how the architecture influences the explanation itself. As a result, it remains uncertain whether the features highlighted by XAI represent disease-relevant signals or architecture-specific biases.

The unexplored question is: how do different neural network architectures (convolutional, recurrent, transformer) influence what the model "sees" in ECG signals - (Figure 1) for AS detection? Convolutional neural networks (CNNs)

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

apply local spatial filters, recurrent networks (LSTMs) capture temporal dependencies, and Vision Transformers (ViTs) employ self-attention across input sequences. These structural differences may lead to divergent attribution maps even on the same ECG dataset, undermining the assumption that XAI outputs are architecture-agnostic. Without systematically comparing explainability across architectures, we cannot know whether AI-ECG tools generalise interpretative validity or whether some models embed hidden biases that mislead clinicians.

To address this gap, our research undertakes a comparative study of CNN, LSTM and ViT-MAE architectures for ECG-based AS screening. We train all models on the same 3-lead ECG dataset, measure diagnostic performance (AUC, sensitivity, specificity), and then apply XAI techniques (Integrated Gradients and LIME) to each model's predictions. By comparing the resulting explanation maps, we assess consistency of feature attribution and identify architecture-dependent biases. Our goal is to deliver practical, architecture-aware interpretability insights that improve clinical trust in AI-ECG screening tools for AS.

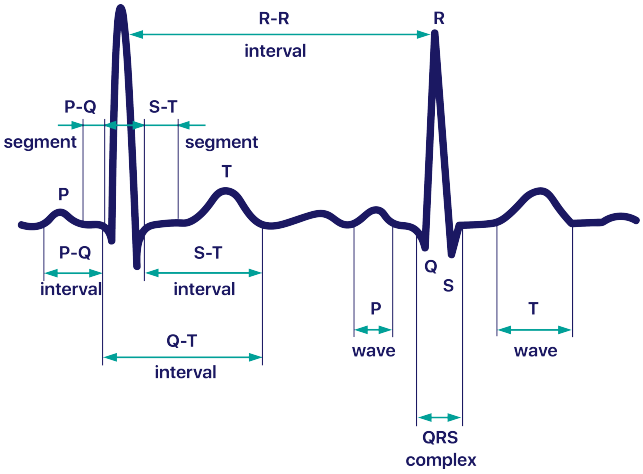

Figure 1: Example ECG characterization plot

Related Work

This section presents existing related work on ECG-based aortic stenosis detection, focusing on classification approaches, architectural innovation, and model explainability. Table 1 summarizes and compares the key studies discussed.

The deployment of deep learning (DL) methods for electrocardiographic (ECG) analysis in aortic stenosis (AS) detection has evolved rapidly, establishing AI-ECG as a viable screening and diagnostic tool. Early studies such as (Purwono, Wulandari, and Nisa 2025) demonstrated the clinical feasibility of convolutional neural networks (CNNs) for AS detection, reporting AUROC values exceeding 0.88 across multicenter cohorts, while more recent works (Francis Densil Raj V 2024; Aminorroaya et al. 2023) extended the task to low-resource and noisy settings using hybrid CNN–RNN–LSTM or single-lead models. These architectures capture both spatial and temporal ECG dependencies, yielding accuracies above 85% even with limited leads. Similarly, (Sanabria et al. 2024) extended this paradigm to longitudinal forecasting of AS progression using tree-based learners. Collectively, these findings affirm that deep learning can reliably identify AS from ECGs, but they also reveal an emerging stagnation—models are optimized for accuracy, yet little attention is given to interpretability or clinical validation.

While classification benchmarks dominate, the interpretability of these architectures remains underexplored, limiting their translational impact. (Purwono, Wulandari, and Nisa 2025) employed gradient-based sensitivity maps to localize CNN attention to precordial T-wave regions but did not examine the complementarity of the MLP and CNN submodules or whether feature localization aligned with cardiological reasoning. (Aminorroaya et al. 2023) provided SHAP-based feature ranking for tabular features but ignored deep feature attribution entirely, leaving unclear whether their noise-adaptive CNN learned physiological or artifact-driven patterns. Likewise, (Francis Densil Raj V 2024) and (Sanabria et al. 2024) focused exclusively on model accuracy, offering no interpretability analysis despite rich multimodal datasets. The consequence is a field where performance metrics are exhaustively compared, but reasoning pathways remain opaque—hindering trust, generalization, and regulatory acceptance.

A more fundamental methodological limitation concerns the architectural isolation of explainability efforts. Existing studies analyze interpretability within individual models but fail to systematically compare how different architectures conceptualize AS-related signal morphology. For instance, (Hata et al. 2020) revealed ST–T segment attention in a VGG16 model using Grad-CAM, yet similar attention localization was found across different lead configurations, suggesting convergent learning not yet explained mechanistically. (Nagai et al. 2025) integrated ECG and chest X-ray data through ResNet–Transformer–EfficientNet pipelines, achieving high AUROC (0.822) but offering no insight into modality or architecture-specific feature attribution. This absence of cross-architectural analysis means researchers cannot discern whether CNNs, Transformers, or recurrent networks rely on shared electrophysiological cues or divergent representational strategies—a crucial gap for model selection, interpretability benchmarking, and ensemble design.

Equally concerning is the near-total absence of clinical validation for model explanations. Few studies assess whether saliency maps or SHAP attributions align with cardiologists' reasoning or echocardiographic gold standards. Even high-performing ensembles such as (Holste et al. 2023)'s 3D-ResNet18 model or (Gan et al. 2025)'s RSMAS-Net report strong predictive accuracy without clinician-grounded verification of model saliency or feature attribution. This gap restricts interpretability to visualization rather than explanation and impedes clinical translation. Therefore, while the field demonstrates that ECG-based deep learning for AS detection is technically feasible, it lacks architecture-aware explainability and clinical validation. The present study addresses these gaps by performing a comparative ex-

plainability analysis across ResNet18, LSTM and ViT-MAE architectures, validating the interpretive consistency of their attributions against expert cardiological judgment—a necessary step toward trustworthy and clinically interpretable AI-ECG diagnostics.

Methodology

In relation to the research objective, the methodology addresses the fundamental question of whether explanations are task-specific or architecture-dependent through systematic multi-architecture comparison using identical datasets and standardized evaluation protocols.

The experimental design follows a three-phase approach:

- Data preprocessing and standardization across all model architectures (ResNet18, LSTM, ViT-MAE ECG).
- Model training and algorithm optimization (RESNet18), LSTM, and ViT-MAE ECG for the binary classification task and
- Comprehensive explainability analysis using the integrated gradients and LIME visualization technique adapted for ECG time-series data.

Our evaluation framework incorporates both quantitative performance metrics (accuracy, precision, recall, F1-score) and novel explainability quality measures that assess clinical relevance and consistency of identified features across architectures.

Data preprocessing and Standardization Across All Model Architectures

1) The Dataset: This study uses an open-access database of cardio-mechanical signals from patients with valvular heart disease, published by (Chenxi et al. 2021). The database uniquely provides synchronized ECG, seismocardiogram (SCG), and gyrocardiogram (GCG) recordings from actual cardiovascular patients, unlike existing databases that only include healthy subjects.The dataset comprises 100 patients from Columbia University Medical Center (New York, USA) and First Affiliated Hospital of Nanjing Medical University (China), with ages ranging from 29-97 years (mean: 68±14) and balanced gender representation (41% female, 59% male). For binary aortic stenosis classification, we utilize 39 patients with moderate or severe AS versus 61 non-AS cases.All recordings were acquired using Shimmer 3 ECG modules with standardized protocols at 256-512 Hz sampling rates. ECG data were captured through a 3-lead limb configuration (LA-RA, LL-LA, LL-RA) with an average recording duration of 6 minutes 48 seconds per patient. The dataset exhibited a substantial class imbalance and limitedness in its original distribution, to address the issue, the study performed data augmentation following the established methodology by (Shiraga et al. 2023)

2) Data Augmentation: The original dataset exhibited a substantial class imbalance, with aortic stenosis (AS) cases occurring far less frequently than non-AS cases, as is typical in clinical datasets. In the original splits, the training set contained 770 AS and 3,257 No-AS samples, while the validation set contained 220 AS and 930 No-AS samples. To address this imbalance and data limitedness, we adopted an oversampling strategy inspired by (Shiraga et al. 2023), applying it exclusively to the training and validation sets while leaving the test set unchanged to ensure unbiased evaluation under realistic clinical conditions.

Oversampling was performed via random sampling with replacement within each class using predefined multiplication factors. The AS class was oversampled by a factor of 21, and the No-AS class by a factor of 5. After oversampling, the training set contained 16,170 AS and 16,285 No-AS samples, while the validation set contained 4,620 AS and 4,650 No-AS samples. This transformation effectively mitigates the original class imbalance and increased the data without introducing synthetic signals, enabling balanced learning during model training while preserving the original class distribution in the test set for robust generalization assessment.

3) Data preprocessing: Electrocardiogram (ECG) signals were subjected to extensive preprocessing to ensure data quality and consistency across all experimental models. Lead standardization was first applied to maintain uniform input dimensions, followed by downsampling to 256 Hz to reduce computational requirements while preserving clinically relevant information. A bandpass filter (0.5-40 Hz) eliminated baseline wander and high-frequency noise, while min-max normalization scaled signal amplitudes to a [0, 1] range for optimal model convergence. This standardized preprocessing pipeline enabled robust feature extraction while maintaining the physiological integrity of the original ECG waveforms.

4) Architecture-Specific Data Formatting: Given the distinct input requirements of each neural network architecture, the preprocessed ECG signals underwent architecture-specific formatting:

1. **CNN Architecture:** The one-dimensional ECG signals (1,280 samples) were reshaped into two-dimensional representations with dimensions $(1 \times 32 \times 40)$, treating the signal as a single-channel image. This transformation enabled the application of 2D convolutional operations while preserving temporal relationships within the ECG morphology.
2. **LSTM Architecture:** ECG signals maintained their sequential nature as one-dimensional time series with shape $(1{,}280 \times 1)$, where each sample represented a temporal feature input to the recurrent network. This format preserved the natural temporal dependencies essential for LSTM processing.
3. **Vision Transformer (ViT) Architecture:** Following the CNN preprocessing, signals were first reshaped to $(1 \times 32 \times 40)$, then upsampled to (224×224) using bilinear interpolation to match standard ViT input dimensions. The single-channel representation was replicated across three channels (RGB format) to create a $(3 \times 224 \times 224)$ input, enabling the use of pre-trained ViT weights.

Hyperparameters and Architecture Configurations for Different Models

1. **ResNet18:** We implemented ResNet18 for ECG classification, using the Adam optimizer with a 1e-3 learning rate and 32 batch size. The CosineAnnealingWarmRestarts scheduler managed learning rate adjustments during training. The model used ResNet18 standard architecture with 18 layer configuration and CrossEntropyLoss for optimization. Additional regularization included a 0.3 dropout and label smoothing of 0.05, with early stopping set at 6 epochs patience to prevent overfitting.
2. **LSTM:** Our LSTM network used AdamW optimizer with a 1e-4 learning rate and 32 batch size. The ReduceLROnPlateau scheduler reduced the learning rate when the validation performance plateaued. The architecture stacked 2 LSTM layers with 128 hidden units, followed by two fully-connected layers, optimized with CrossEntropyLoss. We applied gradient clipping with a maximum norm of 1.0 and early stopping at 7 epochs for stable training.
3. **ViT-MAE ECG:** We utilized a pretrained Vision Transformer Masked Autoencoder (ViT-MAE) that was initially trained on paired ECG and cardiac magnetic resonance (CMR) data from 27,951 subjects in the UK Biobank (Selivanov et al. 2025). This multimodal pretraining aligned ECG signals with CMR-derived functional parameters through contrastive learning, enriching the model's understanding of cardiac physiology beyond electrical activity. For our aortic stenosis detection task, we employed the AdamW optimizer with 1e-4 learning rate and 32 batch size, using CosineAnnealingLR scheduling and CrossEntropyLoss. The architecture featured 12 transformer layers with 12 attention heads, and we applied differential learning rates where the pretrained encoder received 0.1× the base rate while the classifier used the full rate.

Model training and algorithm optimization The architecture outlines an overall workflow for the ECG analysis performed. Initially, a standardized preprocessing pipeline was applied to all raw ECG data to ensure consistency and quality across the dataset. Subsequently, three distinct deep learning architectures—ResNet18, LSTM and a Vision Transformer (ViT-MAE)—were each trained independently on the preprocessed data Both the training and prediction phases were conducted independently for each model to maintain methodological consistency across experiments. After training the models, the Integrated gradients method was uniformly applied across all architectures to interpret their decision-making processes.

Explainability analysis To enhance the interpretability of the proposed models, we applied Integrated Gradients (IG) and Local Interpretable Model-agnostic Explanations (LIME) to analyze the features influencing ECG classification decisions. These complementary explainability methods provide insight into both global attribution patterns and local, instance-specific model behavior.

Integrated gradient To elucidate our model's decision-making, we employ Integrated Gradients (Sundararajan, Taly, and Yan 2017), an attribution method that addresses fundamental limitations of prior approaches. Unlike gradient-based methods that violate Sensitivity (failing to attribute importance to features that actually influence predictions) or backpropagation variants like DeepLift and Layerwise relevance propagation (LRP) that break Implementation Invariance (producing different attributions for functionally equivalent models), IG satisfies both axioms by computing the path integral of gradients along a straight-line path from a neutral baseline to the input. This ensures that attributions reflect the model's actual behavior rather than implementation artifacts. We selected IG for its strong theoretical foundation and consistent performance across architectures. Our implementation uses the Captum library with a zero baseline, computing attributions through 50-step Riemann approximation. The method's Implementation Invariance guarantees reliable explanations within each architecture while accommodating expected differences in attention

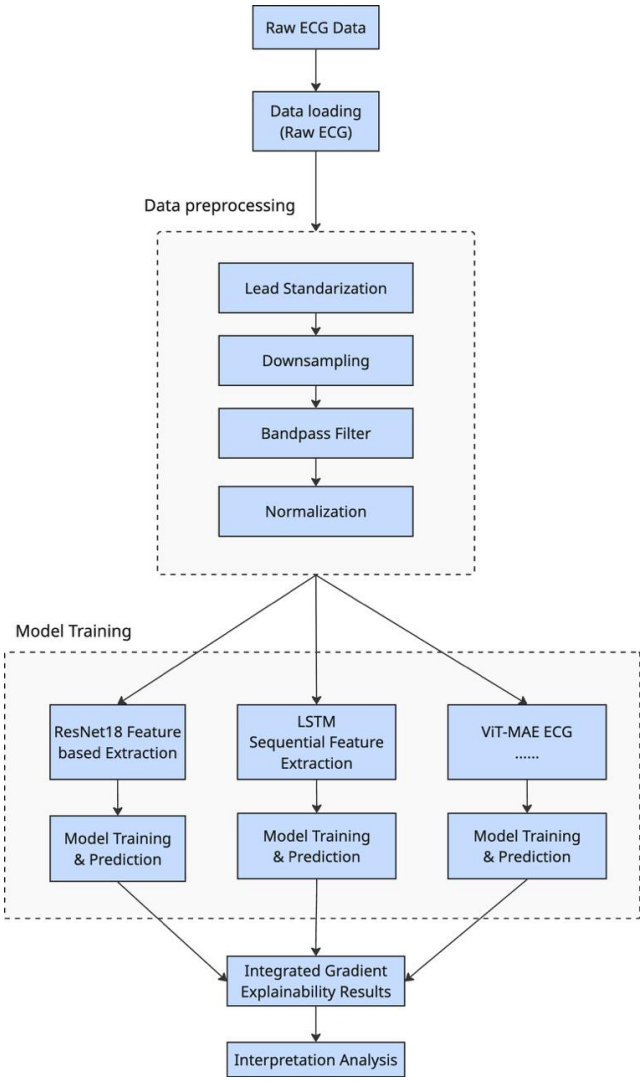

Figure 2: Models Training Overview

patterns across CNN, LSTM, and ViT models due to their distinct inductive biases and learned feature representations.

LIME To complement the gradient-based attributions provided by IG, we used Local Interpretable Model-agnostic Explanations (LIME), a perturbation-based method designed to explain individual model predictions in a locally faithful manner (Ribeiro, Singh, and Guestrin 2016). Unlike IG, which produces continuous attribution scores throughout the input, LIME approximates the behavior of the model in the neighborhood of a specific prediction by learning an interpretable surrogate model from systematically perturbed inputs. This local approximation enables instance-level insight into which regions of the ECG signal most strongly influence a particular classification decision, independent of the underlying model architecture.

In our implementation, the ECG signal is partitioned into a fixed number of contiguous temporal segments, each treated as an interpretable feature. LIME generates perturbed samples by selectively masking these segments,replacing them with segment-wise mean values, and observes the resulting changes in the predicted class probabilities of the model. A sparse linear surrogate model is then fitted to these perturbation-prediction pairs, resulting in segment-level importance weights that reflect the local sensitivity of the classifier. To ensure robustness, we generate several hundred perturbations per explanation and visualize only consistently influential regions after normalization and smoothing.

By mapping the LIME segment-level importance scores back to the original ECG timeline, we obtain localized explanations that can be directly compared with clinically defined waveform components. While LIME does not satisfy axioms such as Implementation Invariance (Sundararajan, Taly, and Yan 2017), its model-agnostic and perturbation-based nature provides complementary insight into decision boundaries and the dependence of local features. When used alongside Integrated Gradients, LIME enables a more comprehensive interpretability analysis by jointly capturing global attribution consistency and local decision sensitivity, thereby revealing architectural biases and instance-specific reasoning patterns across CNN, LSTM, and ViT models.

Results and Discussion

This section presents the performance outcomes of the three models (ResNet18, LSTM, and ViT-MAE). The presentation is divided into three parts: (A) Model classification performance and (B) Explainability findings based on integrated gradient (IG) and LIME analysis. (c) Cross-model comparative analysis.

A. Model Classification Performance: As presented in Tables 1, the classification performance of the three algorithms reveals that the LSTM model demonstrates superior accuracy compared to the others.

B. Explainability Analysis for Models To interpret the decision-making processes of the deep learning models, we employed IG and LIME. IG provides gradient-based attribution maps that quantify the contribution of each input ECG sample to the model's prediction relative to a baseline, offering insight into globally important waveform regions. In contrast, LIME generates local model-agnostic explanations by approximating the behavior of the model in the neighborhood of individual predictions using an interpretable surrogate model. Together, these techniques reveal the specific ECG features prioritized by each architecture during classification, allowing the mapping of relative importance between waveform components and facilitating the assessment of whether the models focus on clinically meaningful patterns. This analysis also exposes architectural biases in the way different models process identical ECG signals.

Prior to model inference, we first identified and demarcated the key components of the cardiac waveform within the original ECG signal—namely P-waves, QRS complexes, T-waves, ST segments, and QT intervals—to establish ground-truth feature locations, as illustrated in Fig. 3. This preliminary signal-level analysis revealed five distinct cardiac beats with clearly delineated waveforms throughout the recording.

Subsequently, the same ECG segment was provided as input to all three model architectures, allowing a controlled comparison of their learned representations. Integrated Gradients were used to generate attribution maps over the full ECG signal, allowing us to assess whether the gradient-based importance scores of the models aligned with the pre-identified diagnostically relevant waveform regions during prediction. In parallel, LIME was applied to individual model predictions to produce local, instance-specific explanations by perturbing segments of the ECG signal and fitting an interpretable surrogate model.

The combined use of IG and LIME enabled a systematic evaluation of both global attribution consistency and local decision sensitivity. This approach allowed us to examine how effectively each architecture attends to known cardiac features and to identify differences in feature utilization and architectural biases across models, as shown in Fig. 2.

1. **ResNeT-18:** The IG method revealed that the CNN architecture demonstrates comprehensive attention to all major ECG waveform components as seen in Fig.4. The attribution maps showed consistent and balanced importance across P-wave, ST segment, QRS complex, and T-wave regions. The model exhibited particularly strong and continuous attribution along the ST segments, suggesting detailed morphological analysis of this clinically critical region. Additionally, the gradient-based importance scores displayed fine-grained patterns within each waveform component, indicating the CNN's capacity for nuanced feature extraction across the entire cardiac cycle without temporal bias.
2. **LSTM:** IG analysis of the LSTM network revealed a distinct temporal processing pattern characterized by primary emphasis on QRS complexes followed by secondary attention to P-wave and ST segment regions(Fig. 5). The attribution maps displayed sequential importance patterns that aligned with the natural temporal progression of cardiac electrical activity. Notably, the method captured the LSTM's specific focus on T-wave offset

Table 1: Comparison of Model Performance Metrics Across ECG-Based Aortic Stenosis Detection Studies

Authors (Year)	Model / Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
(Hata et al. 2020)	VGG16 CNN (12-lead ECG)	79.5	84.2	72.7	78.0
(Hata et al. 2020)	VGG16 CNN (4-lead ECG)	77.3	77.3	77.3	77.3
(Francis Densil Raj V 2024)	CNN-RNN-LSTM	87.0	85.0	86.0	85.5
(Nagai et al. 2025)	ResNet + Transformer + EfficientNet (Multi-modal)	87.9	24.7	49.3	–
Our Work	ViT-MAE ECG (Fine-Tuning)	88.56	97.39	88.20	92.57
Our Work	CNN - ResNet18	97.23	99.34	97.21	98.26
Our Work	LSTM	98.96	98.01	98.67	98.34

This table summarizes accuracy, precision, recall, and F1-score metrics from selected ECG-based aortic stenosis detection studies and compares them with the proposed models (ViT-MAE, LSTM, and CNN) in our work. A dash (–) indicates that the metric was not reported in the respective study.

regions, suggesting the model learned to prioritize the terminal portion of ventricular repolarization. The gradient flows demonstrated how the recurrent architecture maintains and propagates feature importance across time steps, creating interconnected attribution patterns throughout the signal duration.

3. **ViTMAE:** The Vision Transformer exhibited unique attribution characteristics through Integrated Gradients, with predominant focus on QRS complex, T-wave regions (considered as QT interval) and secondary emphasis on P-wave components as shown in Fig. 6. The analysis revealed globally distributed attention patterns that transcended traditional waveform boundaries, showing how the self-attention mechanism processes relationships across different signal segments.The gradient maps displayed patch-wise importance distributions that highlighted the model's ability to capture long-range dependencies and morphological patterns across the entire ECG signal.

LIME Explanations

1. **ResNet-18:** LIME analysis of the CNN architecture indicated a broad and inclusive utilization of ECG waveform components during classification, as illustrated in Fig.8. Locally faithful explanations highlighted consistent contributions from P-waves, QRS complexes, and T-waves across cardiac cycles, with occasional attention to ST segments. These findings suggest that CNN relies on localized morphological features distributed throughout the cardiac cycle, reflecting its strength in capturing spatially contiguous patterns through convolutional operations. Segment-level explanations further indicate that CNN decisions are informed by multiple waveform components rather than a single dominant feature.
2. **LSTM:** The LIME explanations for the LSTM model revealed a decision-making strategy primarily driven by ventricular activity, with strong contributions from QRS complexes and T-wave regions, and secondary involvement of atrial activity represented by P-waves (Fig.9). Localized perturbation analysis suggests that the recurrent architecture places greater emphasis on temporally salient waveform segments that capture dynamic changes throughout the cardiac cycle. This behavior reflects the ability of the LSTM to model temporal dependencies, with the importance of the characteristics concentrated on regions associated with the dynamics of depolarization and repolarization.
3. **ViT-MAE:** For the Vision Transformer, LIME analysis revealed a dominant reliance on QRS complexes and repolarization related regions, particularly T-waves and ST segments, as shown in Fig.10. The explanations exhibited distributed yet selective sensitivity to key diagnostic regions, consistent with the transformer's patch-based self-attention mechanism. Unlike the CNN and LSTM, the ViT demonstrated a tendency to aggregate information across multiple waveform segments, indicating a more global interpretation of ECG morphology while still prioritizing clinically relevant features.

C. Cross-Model Comparative Analysis for IG :

The comparative analysis of Integrated Gradients attributions reveals fundamental differences in how each architecture processes ECG signals, reflecting their underlying structural biases and learning mechanisms. The CNN demonstrates the most comprehensive approach, with balanced attention distributed across all major waveform components—P-wave, ST segment, QRS complex, and T-wave—suggesting it develops a holistic understanding of cardiac electrical activity similar to human expert analysis. In contrast, the LSTM exhibits distinct temporal prioritization, emphasizing QRS complexes as primary features while maintaining secondary focus on P-wave and ST segments, indicating its recurrent architecture leverages sequential dependencies between depolarization and repolarization events. Most strikingly, the Vision Transformer shows a unique global pattern recognition strategy, prioritizing T-wave and P-wave regions with reduced QRS complex attention, suggesting that its self-attention mechanism captures different morphological relationships than the convolutional or recurrent approaches. These architectural differences manifest clearly through Integrated Gradients, with CNN showing robust feature utilization across all clinically relevant components, LSTM demonstrating time-dependent

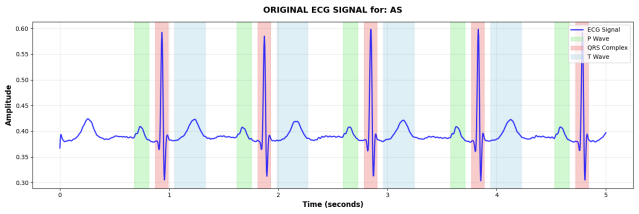

Figure 3: Original ECG Signal with Features Detected.

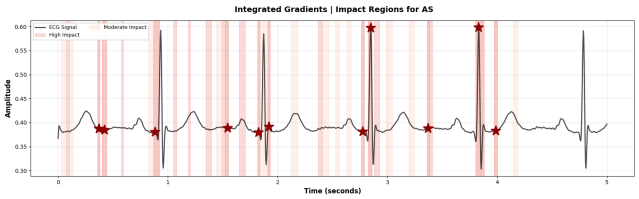

Figure 4: Integrated Gradients – CNN Model

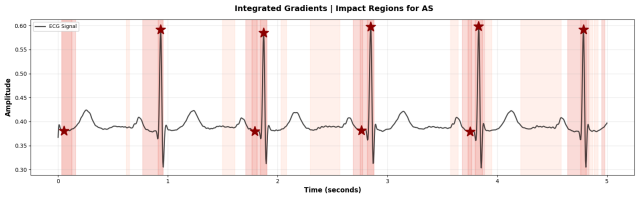

Figure 5: Integrated Gradients – LSTM Model

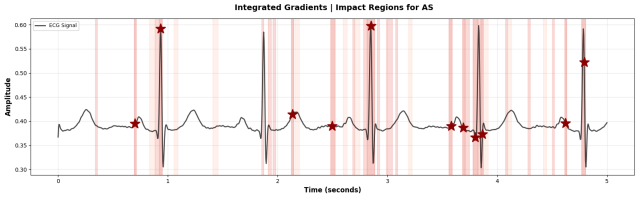

Figure 6: Integrated Gradients – ViT Model

Figure 7: Comparison of Integrated Gradient Visualizations across CNN, LSTM, and ViT Models relative to the original ECG waveform. Each subfigure highlights how different architectures attend to various cardiac waveform components.

processing hierarchies, and ViT revealing unconventional but potentially complementary pattern recognition strategies that emphasize repolarization characteristics over depolarization events.

D. Cross-Model Comparative Analysis for LIME The comparative analysis of LIME explanations highlights clear differences in how each architecture relies on localized ECG regions when forming individual predictions. CNN exhibits a distributed dependence on multiple waveform components, with consistent sensitivity to P-waves, QRS complexes, and T-waves, suggesting that its decisions are guided by a combination of local morphological cues throughout the cardiac cycle. The LSTM, in contrast, shows a more selective pattern, with explanations dominated by ventricular activity, particularly QRS complexes and T-waves, reflecting its emphasis on temporally salient segments that carry

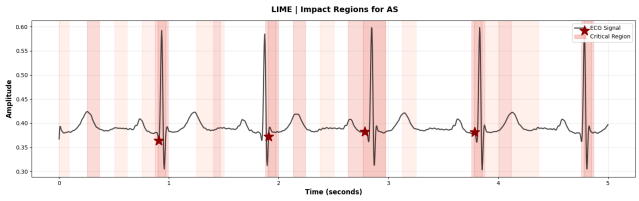

Figure 8: LIME – CNN Model

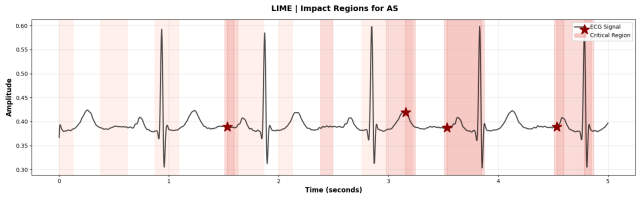

Figure 9: LIME – LSTM Model

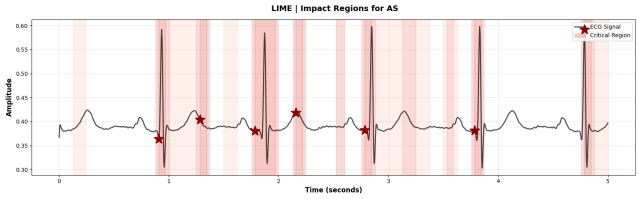

Figure 10: LIME – ViT Model

Figure 11: Comparison of LIME Visualizations across CNN, LSTM, and ViT Models relative to the original ECG waveform. Each subfigure highlights how different architectures attend to various cardiac waveform components.

sequential information. The Vision Transformer displays a distinct behavior, with strong reliance on QRS complexes and frequent involvement of repolarization-related regions such as T-waves and ST segments, indicating a tendency to aggregate information across broader signal segments rather than focusing on isolated events. These differences underline how LIME captures architecture-specific decision strategies at the instance level, revealing how each model locally responds to perturbations in clinically relevant ECG regions.

E. LIME and IG Comparative Analysis A comparison of Integrated Gradients and LIME highlights how different architectures rely on distinct ECG features when making predictions. Integrated Gradients shows that the CNN spreads attention across all major waveform components, while the LSTM concentrates more strongly on temporally salient regions such as the QRS complex and T-wave. The Vision Transformer, in contrast, places greater emphasis on repolarization-related regions and distributes the importance more globally across the signal. LIME supports these observations at the instance level. For example, CNN explanations consistently involve multiple waveform segments within a cardiac cycle, whereas LSTM explanations are dominated by ventricular activity, and Vision Transformer explanations frequently highlight both depolarization and ST–T regions.

Although the two methods operate differently, they point to the same pattern: each architecture favors different ECG features. Taken together, these results demonstrate that the explanation outcomes are influenced by the model architecture, as the features identified as important vary systematically across CNN, LSTM, and Vision Transformer models.

Conclusion

This study demonstrates that neural network architecture plays a decisive role in shaping explainability outcomes for ECG-based aortic stenosis detection, even when models achieve comparable predictive performance. By jointly applying IG and LIME to ResNet18, LSTM, and ViT-MAE architectures, we show that each model consistently relies on different components of the ECG waveform during decision-making. Integrated Gradients reveal stable, architecture-specific attribution patterns across the full signal, while LIME confirms these differences at the instance level by highlighting distinct locally influential regions.

Across both explainability methods, CNN exhibits a broad attention to multiple waveform components, indicating a holistic processing strategy. The LSTM places a stronger emphasis on temporally salient regions, particularly QRS complexes and T-waves, reflecting its sequential modeling bias. In contrast, the Vision Transformer relies more heavily on repolarization related regions and distributed signal segments, consistent with its global self-attention mechanism. The agreement between Integrated Gradients and LIME strengthens the conclusion that these differences arise from architectural inductive biases rather than artifacts of a single explainability technique.

Importantly, both methods indicate that all architectures attend to clinically meaningful ECG regions established in cardiovascular literature, including QRS complexes, ST segments, ST–T waves, and QT intervals. While the specific regions emphasized vary by architecture, this convergence on medically relevant features supports the physiological plausibility of the learned representations. For example, CNN sensitivity to ST segments and Vision Transformer's attention to QT-related regions align with known electrophysiological manifestations of aortic stenosis.

These findings have direct implications for the development of trustworthy AI in cardiovascular medicine. Our results suggest that explainability should not be treated as architecture agnostic, as different model designs yield systematically different interpretations. Instead, architecture-aware explainability is necessary to ensure that model explanations remain consistent, clinically meaningful, and interpretable to medical practitioners. By combining global and local explanation methods, this work contributes to the development of transparent and reliable AI systems for ECG analysis that can better support clinical decision-making and foster trust in real-world deployment.

References

Aminorroaya, A.; Dhingra, L. S.; Sangha, V.; Oikonomou, E. K.; Khunte, A.; Shankar, S. V.; Camargos, A. P.; Haynes, N. A.; Hofer, I.; Ouyang, D.; Nadkarni, G. N.; and Khera, R. 2023. Deep Learning-enabled Detection of Aortic Stenosis from Noisy Single Lead Electrocardiograms.

Chenxi, Y.; Foli, F.; Nicole, A.; Philip, G.; Yuwen, L.; Liu, C.; and Tavassolian, N. 2021. An Open-access Database for the Evaluation of Cardio-mechanical Signals from Patients with Valvular Heart Diseases.

Francis Densil Raj V. 2024. A Novel CNN-RNN-LSTM Framework for Predictive Cardiovascular Diagnostics of Aortic Stenosis in a Large Scale 12-Lead ECG Dataset. *Communications on Applied Nonlinear Analysis*, 32(3): 685–700.

Gan, Y.; Huang, W.; Deng, Y.; Xie, X.; Gu, Y.; Zhou, Y.; Zhang, Q.; Zhang, M.; and Liu, Y. 2025. RAMAS-Net: a module-optimized convolutional network model for aortic valve stenosis recognition in echocardiography. *Frontiers in Medicine*, 12: 1587307.

GBD Cardiovascular Disease Collaboration. 2023. Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021. *Journal of the American College of Cardiology*.

Hata, E.; Seo, C.; Nakayama, M.; Iwasaki, K.; Ohkawauchi, T.; and Ohya, J. 2020. Classification of Aortic Stenosis Using ECG by Deep Learning and its Analysis Using Grad-CAM. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1548–1551. Montreal, QC, Canada: IEEE. ISBN 978-1-7281-1990-8.

Holste, G.; Oikonomou, E. K.; Mortazavi, B. J.; Coppi, A.; Faridi, K. F.; Miller, E. J.; Forrest, J. K.; McNamara, R. L.; Ohno-Machado, L.; Yuan, N.; Gupta, A.; Ouyang, D.; Krumholz, H. M.; Wang, Z.; and Khera, R. 2023. Severe aortic stenosis detection by deep learning applied to echocardiography. *European Heart Journal*, 44(43): 4592–4604.

Institute for Health Metrics and Evaluation. 2023. Report: cardiovascular diseases caused 1 in 3 global deaths in 2023.

Iung, B.; Baron, G.; Butchart, E. G.; Delahaye, F.; Gohlke-Bärwolf, C.; Levang, O. W.; Tornos, P.; Vanoverschelde, J.-L.; Vermeer, F.; Boersma, E.; Ravaud, P.; and Vahanian, A. 2003. A prospective survey of patients with valvular heart disease in Europe: The Euro Heart Survey on Valvular Heart Disease. *European Heart Journal*, 24(13): 1231–1243. _eprint: https://academic.oup.com/eurheartj/article-pdf/24/13/1231/17885967/1231.pdf.

Kwon, J.; Lee, S. Y.; Jeon, K.; Lee, Y.; Kim, K.; Park, J.; Oh, B.; and Lee, M. 2020. Deep Learning–Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. *Journal of the American Heart Association*, 9(7): e014717.

Nagai, S.; Nishimori, M.; Shinohara, M.; and Tanaka, H. 2025. Enhanced detection of aortic stenosis using a multimodal deep learning approach combining ECG and chest X-ray data. *European Heart Journal - Cardiovascular Imaging*, 26(Supplement_1): jeae333.424.

Purwono, P.; Wulandari, A. N. E.; and Nisa, K. 2025. Explainable Artificial Intelligence (XAI) in Medical Imaging: Techniques, Applications, Challenges, and Future Directions.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any

Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Sanabria, M.; Tastet, L.; Pelletier, S.; Leclercq, M.; Ohl, L.; Hermann, L.; Mattei, P.-A.; Precioso, F.; Coté, N.; Pibarot, P.; and Droit, A. 2024. AI-Enhanced Prediction of Aortic Stenosis Progression. *JACC: Advances*, 3(10): 101234.

Selivanov, A.; Müller, P.; Özgün Turgut; Stolt-Ansó, N.; and Rückert, D. 2025. Global and Local Contrastive Learning for Joint Representations from Cardiac MRI and ECG. arXiv:2506.20683.

Shiraga, T.; Makimoto, H.; Kohlmann, B.; Magnisali, C.-E.; Imai, Y.; Itani, Y.; Makimoto, A.; Schölzel, F.; Bejinariu, A.; Kelm, M.; and Rana, O. 2023. Improving Valvular Pathologies and Ventricular Dysfunction Diagnostic Efficiency Using Combined Auscultation and Electrocardiography Data: A Multimodal AI Approach. *Sensors*, 23(24): 9834.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365.

Thoenes, M.; Bramlage, P.; Zamorano, P.; Messika-Zeitoun, D.; Wendt, D.; Kasel, M.; Kurucova, J.; and Steeds, R. P. 2018. Patient screening for early detection of aortic stenosis (AS)—review of current practice and future perspectives. *Journal of Thoracic Disease*, 10(9). Publisher: AME Publishing Company.

van Geldorp, M. W.; van Gameren, M.; Kappetein, A. P.; Arabkhani, B.; de Groot-de Laat, L. E.; Takkenberg, J. J.; and Bogers, A. J. 2009. Therapeutic decisions for patients with symptomatic severe aortic stenosis: room for improvement?. *European Journal of Cardio-Thoracic Surgery*, 35(6): 953–957. _eprint: https://academic.oup.com/ejcts/article-pdf/35/6/953/17772307/35-6-953.pdf.

WHO. 2025. Cardiovascular diseases.