Evaluating the Role of Data and Explanation Representations in Clinical XAI for Conduction Disorder Classification

Karol Dobiczek[1,2,*,†], Maciej Mozolewski[1,†], Szymon Bobek[1], Michał Szafarczyk[3], Peter van Dam[3] and Grzegorz J. Nalepa[1]

[1]Department of Human-Centered Artificial Intelligence, Institute of Applied Computer Science, Jagiellonian University, Krakow, Poland
[2]Jagiellonian Human-Centered AI Lab, Mark Kac Center for Complex Systems Research, Jagiellonian University, Kraków, Poland
[3]Faculty of Medicine, Jagiellonian University Medical College, Krakow, Poland

Abstract

Clinical XAI depends not only on the explanation method, but also on the representation through which the model's evidence is presented to experts. Through the conduction disorder classification task, we study this dependence on explanation and data representation, using conventional 12-lead ECG and a 3D ventricular trajectory representation of the ECG called CineECG, and two types of model explanations. We perform evaluations at the model level, using classification performance, attribution analysis, and CineECG axis-wise ablation, and at the user level through an expert study comparing Integrated Gradients attributions with Post-hoc Attribution Rules validity corridors. Our 12-lead ECG models achieved higher performance than CineECG models, and experts primarily relied on the standard ECG view even when CineECG was available. The two explanation formats received similar overall ratings, where validity corridors were slightly favored for clinical agreement and informativeness, while attribution heatmaps were slightly favored for interpretability. Our results suggest that CineECG can be treated as a complementary representation and that clinical XAI interfaces should combine familiar diagnostic views with complementary explanation formats.

Keywords

Explainable AI, Deep Learning, Conduction Disorders, CineECG, Feature Attributions, Rule-Based Explanations

CEUR-WS.org/Vol-4339/paper_2.pdf

1. Introduction

The near-perfect performance of deep learning (DL) artificial intelligence (AI) models in cardiac disorder classification and their implementations in tasks such as early detection or decision support has sparked research in explaining these predictive models. Although the complexity and size of these models hinder the ease of interpretation, explainable AI (XAI) methods offer tools aiming to reveal the reasoning behind the model's predictions. Evaluating the accuracy of these methods has been the focus of numerous works [1, 2, 3], evaluating XAI methods along metrics like accuracy or faithfulness. While these are important factors in the development and implementation of XAI methods, it is also crucial to consider the cognitive load that the use and interpretation of XAI imposes on the users. Previous work has shown that the way a model explanation is presented to the system's final users can affect the understandability of the explanations, and thus decrease the confidence in the AI system [4].

In a classical setting, the 12-lead representation serves as the main data source for the clinician's diagnosis. When working with model explanations, either as a decision support tool or verifying its prediction, the clinician must consider both the explanation and the ECG signals at the same time. This

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 July 2026, Bremen, Germany

*Corresponding author.
†These authors contributed equally.
karol.dobiczek@uj.edu.pl (K. Dobiczek); m.mozolewski@uj.edu.pl (M. Mozolewski); szymon.bobek@uj.edu.pl (S. Bobek); michal.szafarczyk@student.uj.edu.pl (M. Szafarczyk); peter.dam@uj.edu.pl (P. v. Dam); grzegorz.j.nalepa@uj.edu.pl (G. J. Nalepa)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

setting makes it possible for both the rich ECG data presentation and the explanation type to negatively affect the interpretation.

A recent advancement in cardiac diagnostics called CineECG [5] aims to alleviate the cognitive burden of 12-lead ECG interpretation by reconstructing the propagation pathway of the cardiac electrical signal, mapping it directly from ECG data. Furthermore, this tool provides an alternative perspective on the heart's electrical pathway, which can serve as a supplementary modality to verify and assess standard 12-lead-based diagnoses. The mapping allows for relating the heart signal to the anatomy of the human body, rather than relying on the more abstract projection captured at the 12 leads. Due to this fact, it is theorized that the 3D representation of the signal might remove confounding factors in the data and make it easier to diagnose pathologies.

In this work, we aim to quantify the effect of the data and explanation representations on the task of analyzing ECG model explanations. To this end, we employ classifier models trained on a cardiac pathology prediction task using 12-lead ECG and CineECG representations. We generate explanations for these models using two distinct explainer types: an attribution map and a rule-based explainer. The generated explanations are then presented to clinicians and evaluated under the following criteria: understandability, clinical agreement, and impact on decision-making. Additionally, we perform a comparative analysis of the attributions and model performances of the two data modalities.

The main contributions of our work are:

- A study focused on a single type of disorder with explanations verified by expert clinicians. This focused analysis allows us to pinpoint with higher accuracy the ways in which an explanation aligns with the clinical knowledge. By analyzing single-heartbeat samples, the models, explanations, and evaluators can be focused on the features of a single heart cycle, rather than the differences and anomalies between them.
- An analysis of both the data and attribution representation on the usefulness of explanations for domain experts working with ML models.

This article is organized as follows. Section 2 summarizes related work. Section 3 introduces the data and explanation representations used in the study. Section 4 describes the expert evaluation protocol, while Section 5 presents the dataset, preprocessing, and model-training setup. Section 6 reports the model-level and expert-evaluation results. Section 7 discusses the main findings and limitations, and Section 8 concludes the paper.

2. Related Works

As an established task in machine learning, DL ECG classifiers have been thoroughly studied as a target for the use of XAI methodologies in the past. Most applications focus on explaining these models using post-hoc attribution methods such as SHAP values [6, 7, 8], LIME [7, 8], Grad-CAM [9, 8], Others, instead, aim to establish trustworthiness in the models through the use of explainable-by-design models [10]. This increased popularity has also resulted in works that focus on evaluating the attributions along several metrics, mostly using ECG data as part of a benchmark, not as a standalone task [11, 12, 13].

Until now, studies that focus on evaluating the explainability in ECG have focused on assessment and selection of optimal methods for classification tasks that involve multiple heart cycle ECG data [1]. A recent work involved the analysis of the alignment of model attributions to expert clinician annotations using two data ECG data representations: 12-lead ECG and CineECG [14]. We use the aforementioned work and its dual data representation setup as a starting point for our work.

While the representation or format of an explanation is shown to be one of the most important factors in XAI [15, 4], it remains largely unexplored in the medical domain. This issue is especially relevant for time-series explanations, where attribution maps indicate important time points or channels but do not directly express value constraints on the signal. PHAR (Post-hoc Attribution Rules) addresses this limitation by transforming post-hoc attributions into interval-based rules visualized as validity corridors [16]. Related ECG work has also used attribution-derived intervals and prototype-based

counterfactuals to generate sparse explanations for multivariate 12-lead ECG classifiers [17]. In contrast, our study focuses on how clinical experts use different data and explanation representations when conventional ECG and CineECG are presented together.

3. Preliminaries

In this section, we introduce the necessary aspects of our evaluation scheme. The types of explanations used are defined, and the motivation behind their use is stated. Finally, we introduce the two data representation types.

3.1. Model Attributions

Model attribution methods form one of the main branches of explainable AI. The base role of model attributions is to highlight the parts of the model input that affect its output the most. Most of these methods provide explanations as a single value per feature denoting a negative or positive influence on the target class.

Attribution methods can roughly be divided into two categories: model-agnostic and model-specific methods. Model agnostic methods can be applied to any model, while the latter might use the model architecture and its operation graph to relate the output to the input importance. One such method called Integrated Gradients (IG) [18] has shown the best adherence to the ground truth annotations in the task of single heart cycle classification [14]. Integrated Gradients computes a path-integrated gradient between the explained sample and a baseline sample. The method is grounded in two base axioms: sensitivity – if a change in one feature changes the model output, that feature should have a non-zero attribution, and implementation invariance – attributions should be the same for two functionally identical networks.

The visualization of feature attributions depends on the model architecture; however, they are usually presented as per-timestep attribution scores. Figure 1 presents an example IG heatmap visualization from our experimental setup. While the feature attribution approach to explainability in the time domain is relatively easy to interpret, the relation of attributions to the signal values might be less obvious, thus limiting the explanatory power of the methods. Other methods have been developed with further capabilities in mind, such as feature interactions, modality interactions, or rule-based explanations.

Rule-based explanations express model behavior as explicit, human-inspectable conditions rather than only as relevance scores. For temporal signals, this is useful because experts often reason about waveform morphology, temporal segments, and admissible value ranges. In this work, we use Post-hoc Attribution Rules (PHAR), which transform numeric post-hoc attributions into interval-based rules for time-series data [16]. PHAR selects relevant time–channel features from attribution scores and estimates semi-factual value intervals within which the original model prediction is preserved. The rules derived from those intervals constrain only the selected time–channel features, while the remaining parts of the signal are left unconstrained.

For ECG signals, these rules can be visualized as validity corridors, showing both *where* the relevant signal conditions occur and *what* amplitude ranges remain compatible with the prediction. Figure 1 illustrates this representation in the evaluation view: PHAR conditions are shown as red corridors overlaid on the CineECG and 12-lead ECG signals, alongside the attribution heatmaps used to localize model-relevant regions.

3.2. Conduction Disorders

Conduction disorders (CD) in the human heart stem from pathologies that cause the heart signal to bypass the usual route and propagate through the right or left part of the heart. This phenomenon is observed on an electrocardiogram (ECG) as characteristic features in select leads within the QRS part of the heart cycle.

Conduction disorders form a characteristic group of conditions, including incomplete or complete left or right bundle branch blocks, left anterior or posterior fascicular blocks, and left or right ventricular hypertrophies. Such cases are of particular interest in our study because they present distinct morphological changes in both CineECG and standard 12-lead ECG. This choice of conditions is motivated by the aim to further decrease the cognitive load of the expert annotators by letting them focus on a small set of possible heart conditions.

To train our models for these conditions, we use the publicly available PTB-XL dataset, which contains 10-second ECG signals along with their respective rates, durations, and diagnoses. We segment the ECG signals into 400-ms single-cycle samples. Finally, we map these 12-lead ECG samples to a 3D representation using CineECG.

4. Expert evaluation

To evaluate the impact of the representations on the utility of the generated explanations, we conducted a within-subject expert evaluation involving physicians and cardiologists. The evaluation was carried out via a custom web-based platform where each participant received a unique access token to ensure a controlled and blinded review process.

Each expert evaluated eight clinical cases. In every case, the expert was shown both data representations: the standard 12-lead ECG waveform and the corresponding CineECG ventricular trajectory. This joint presentation was used intentionally because CineECG is a relatively novel representation and was expected to be interpreted alongside the clinically familiar 12-lead ECG rather than in isolation. The experts' reliance on either representation was measured explicitly using a modality-reliance question.

The explanation condition was varied across trials. Each expert was presented with four cases with PHAR-derived validity corridors and four cases with Integrated Gradients attributions. The assignment of explanation type to cases was balanced within each expert session and rotated across experts. The four complete expert sessions yielded 32 trial-level evaluations. Across these completed sessions, each clinical case was evaluated under both explanation conditions at least once.

4.1. Data Representations

Each trial presented both data representations jointly. Both modalities were presented as temporal plots, while CineECG was additionally displayed as a 3D ventricular trajectory. This setup allowed experts to inspect the novel CineECG representation while retaining access to the clinically established ECG view. In the annotation form, experts were then asked which modality they primarily relied on during the evaluation.

4.2. Explanation Types

Each trial was assigned to one of two explanation conditions:

1. **Integrated Gradients heatmaps:** The expert evaluated the 12-lead ECG and CineECG representations overlaid with feature attribution heatmaps from Integrated Gradients attributions. These visualizations indicate where the model assigns importance in the input but do not specify explicit value constraints on the signal.
2. **PHAR validity corridors:** The expert evaluated semi-factual intervals generated by the PHAR method. These explanations were presented as verifiable "validity corridors" superimposed on the signal, explicitly specifying amplitude ranges and temporal segments within which the model prediction is preserved.

Figure 1 illustrates the visual presentation used in the evaluation. Both data representations were shown in each trial: the CineECG representation as a 3D trajectory with per-axis plots, and the standard ECG representation as a two-column grid separating limb and precordial leads. In the Integrated Gradients condition, feature attribution heatmaps were overlaid on the displayed signals to indicate

	Clinical case								Σ	
Expert	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**P**	**IG**
E1	P	P	P	P	IG	IG	IG	IG	4	4
E2	IG	IG	IG	IG	P	P	P	P	4	4
E3	IG	IG	P	P	IG	IG	P	P	4	4
E4	P	IG	P	IG	IG	P	IG	P	4	4
ΣP	**2**	**1**	**3**	**2**	**1**	**2**	**2**	**3**	-	-
ΣIG	**2**	**3**	**1**	**2**	**3**	**2**	**2**	**1**	-	-

Table 1
Assignment of explanation conditions in the four complete expert evaluation sessions. Each expert evaluated four PHAR trials and four Integrated Gradients trials. Legend: P – PHAR validity corridors, IG – Integrated Gradients heatmaps.

where the model assigned importance. In the PHAR condition, validity corridors were used to indicate *what* amplitude constraints had to be preserved over specific temporal segments to maintain the model prediction.

This setup ensured that each expert evaluated both explanation types while always having access to both data representations. Consequently, the study compares the perceived utility of PHAR validity corridors and Integrated Gradients heatmaps, while modality reliance is analyzed as an expert-reported outcome.

4.3. Annotation Questions

After reviewing each case, experts quantified the quality of the explanations using a 5-point Likert scale across three key dimensions:

1. **Understandability:** "How easy is this explanation to interpret?"
   Scale: 1: Very difficult, 2: Difficult, 3: Neutral, 4: Easy, 5: Very easy
2. **Clinical Agreement:** "To what extent does the explanation overlap with your medical knowledge?"
   Scale: 1: Very inaccurate (does not highlight important features, highlights unimportant features), 2: Inaccurate, 3: Neutral, 4: Accurate, 5: Very accurate (highlights clinically important features)
3. **Impact on Decision:** "Does this explanation increase your confidence in the model's diagnosis?"
   Scale: 1: Very unhelpful (distracts from the important features), 2: Unhelpful, 3: Neutral (neither helpful nor unhelpful), 4: Helpful, 5: Very helpful (helps focus on clinically important parts of the data)

The additional guidelines for the latter two questions (in parentheses) are added to ensure that each scale point measures similar feedback among the experts.

We quantify the dependence of experts on each data representation introduced in Section 4.1 by asking an additional question to specify *"Which data modality did [they] primarily rely on for this evaluation"* with five possible answers: *CineECG Only*, *Mostly CineECG*, *Balanced*, *Mostly 12-lead*, *12-lead Only*.

Finally, experts can provide their clinical notes as a free-text input. Crucially, they are asked to specify the exact timeframe(s) in milliseconds (e.g., 50-100 ms, 300-350 ms) that were most critical for their evaluation. These additional notes serve as a means of evaluating the exact parts of the inputs salient for each of the three data classes.

5. Experimental Setup

In this section, we describe the data and models used in our experiments. We state the data preprocessing steps as well as the model selection and tuning procedures.

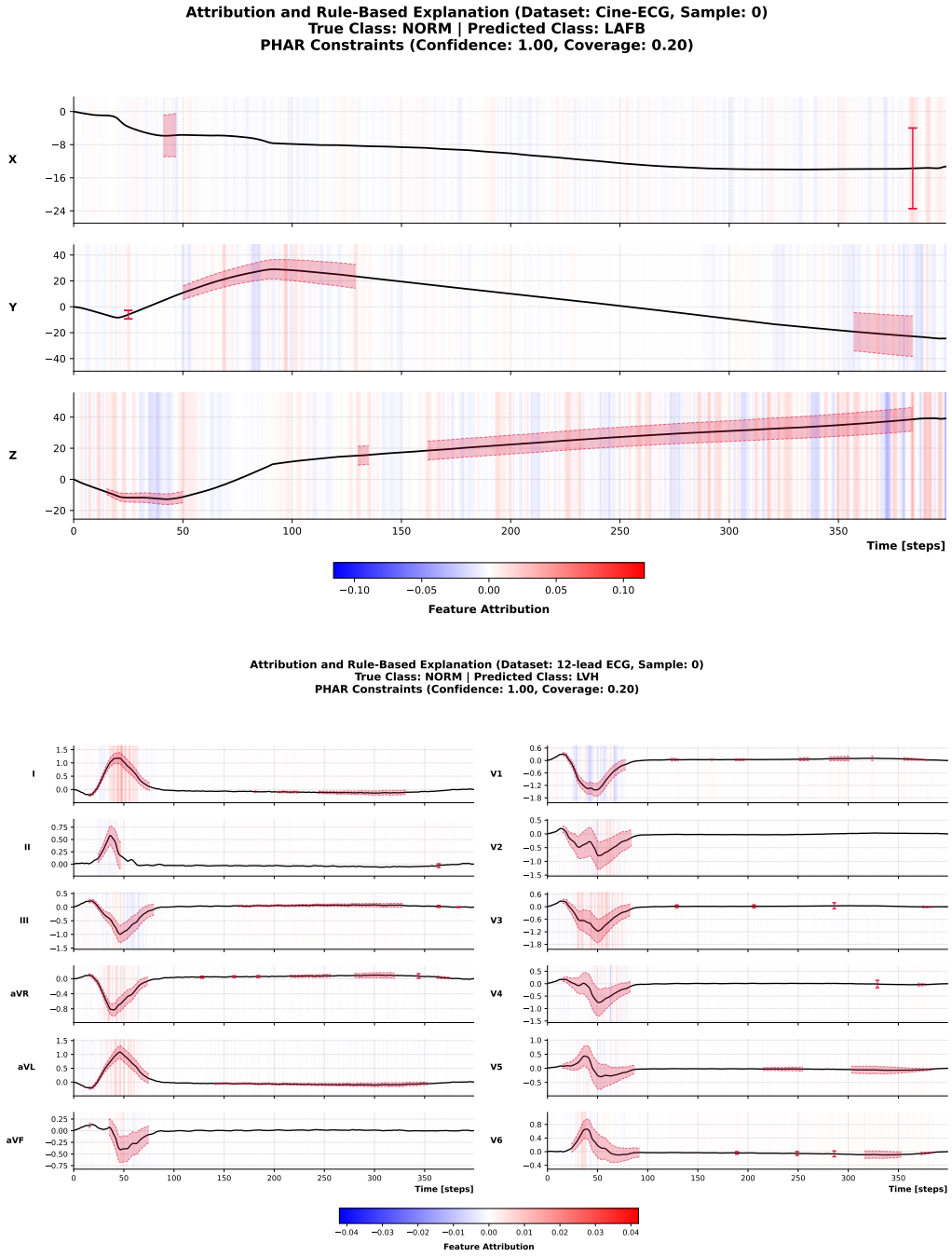


Figure 1: Combined XAI profiles for the same clinical observation (Sample 0). Top: 3-channel CineECG with a feature attribution colormap. Bottom: Standard 12-lead ECG with a feature attribution colormap. In both representations, PHAR interval rules (red corridors) establish strict amplitude constraints over the continuous attribution heatmaps.

5.1. Data

We use the PTB-XL database [19] of ECG signals as the dataset for our experiments and evaluations. The base dataset contains over 20,000 12-lead ECG cases of 10-second length. Each case is labeled with an SCP-ECG statement containing several disorder codes along with their corresponding annotator certainty.

We pre-process the data by retaining only the first sample for each patient. We split the data into classes by selecting the SCP code with the highest annotator certainty. Each 400-millisecond 12-lead data sample is transformed into the 3D ventricular representation using the CineECG framework.

To let the expert annotators focus on a single type of disorders and to test the predictive strength of CineECG, we limit our experiments to the conduction disorder set of classes.

5.1.1. Data Split

For every experiment, we also maintain several healthy samples as an additional baseline class. In our sample selection, for each class, we select patients that were diagnosed with at least one of the CDs, regardless of whether they have any other diagnoses. Ties between CDs are solved based on the label confidence in the dataset.

We study two classification tasks using this dataset: 1. A per-class classification, and 2. a more restrained 3-class normal/left/right task. The per-class classification contains 7 classes: NORM, LVH, LAFB, CLBBB, IRBBB, CRBBB, IVCD. To fight class imbalance and achieve a reasonable F1 score for the model in the multiclass classification scenario, we remove the LPFB, ILBBB, and RVH classes from our data, as they comprise less than one percent of the data. The performance of the model matters in our setting, as the attributions of a poorly trained model might not be meaningful and thus distract from the task. To capture only the differences between the conditions focused on the left and right sides of the heart, we form the second task. We split the data into three classes: *normal*, *left* consisting of LVH, LAFB, LPFB, ILBBB, CLBBB, and *right* consisting of IRBBB, CRBBB, RVH.

We split the data into training, validation, and test sets. The class distribution for the 7-class classification task is shown in Table 2. The validation and test sets have identical class distributions, with 931 samples each.

Class	Diagnosis	Train	Validation	Test
0	NORM	600 (21.5%)	200 (21.5%)	200 (21.5%)
1	LVH	600 (21.5%)	200 (21.5%)	200 (21.5%)
2	LAFB	599 (21.4%)	200 (21.5%)	200 (21.5%)
3	CLBBB	167 (6.0%)	56 (6.0%)	56 (6.0%)
4	IRBBB	383 (13.7%)	127 (13.6%)	127 (13.6%)
5	CRBBB	210 (7.5%)	70 (7.5%)	70 (7.5%)
6	IVCD	234 (8.4%)	78 (8.4%)	78 (8.4%)
Total		**2793**	**931**	**931**

Table 2
Class distribution for the 7-class conduction disorder classification task.

In the 3-class case, we split the samples evenly for each class with 600 samples per class for the train set (total of 1800 samples), and 200 samples per class for the validation and test sets (total of 600 samples per set.)

5.2. Model

The ResNet architecture stands as a strong and established model for classifying ECG data [1]. Similar to [14], we utilize a customized 1D-ResNet model. The architecture, illustrated in Figure 2, is initialized with a strided convolution rather than a standard pooling operation. This change aims to produce less fragmented attributions when using gradient-based methods. Following that, the stacked residual blocks with replication padding aim to mitigate the discontinuities at the channel boundaries. Finally, a Global Average Pooling layer aggregates the multi-channel features into a single latent vector that is regularized through dropout before the classification. This allows us to use the same architecture for the classification of both the 12-lead ECG data and the 3-dimensional CineECG data.

For each input modality, we perform hyperparameter optimization using the Optuna Python library. We optimize the number of filters in the base convolutional layer, the number of ResNet blocks, and training parameters, learning rate, weight decay, and batch size. The hyperparameter optimization is

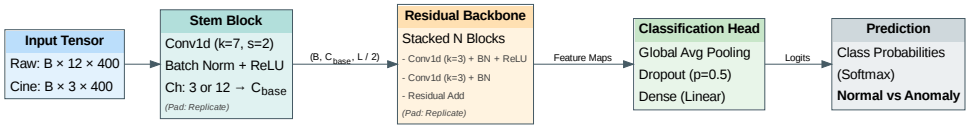


Figure 2: 1D-ResNet architecture used in our experiments.

executed against a separate validation set on which we calculate the accuracy and F-scores. Optuna optimizes the latter.

To further counteract the class imbalance mentioned in Section 5.1.1, we apply class weighing in the cross-entropy loss criterion. The class weights are computed as the inverse of the number of training instances per class.

6. Results

This section presents the results of our experiments. First, we present the performance of the models on the CD data, along with the XAI analysis of these models with regard to different data classes. Second, we show the preferences of the expert annotators in the data presentation and explanation type dimensions.

6.1. Model Performance

The 1D-ResNet models are optimized using Optuna for 100 trials, after which the best-performing model configuration is trained with early stopping based on the validation score.

Task	Data type	F1-score	Accuracy
7-class	12-lead	0.76	0.77
	CineECG	0.64	0.65
3-class	12-lead	0.82	0.83
	CineECG	0.76	0.77

Table 3
Training results for the 7-class and 3-class tasks on the 12-lead and CineECG data.

Table 3 presents the F1-scores and the accuracies for the models trained on both tasks and both data modalities. For both classification tasks, the 12-lead ECG representation models perform better than their CineECG counterparts, showing a difference in both F1 score and accuracy of 12 and 6 percentage points for the 7-class and 3-class tasks, respectively. Additionally, for both data modalities, the 3-class models perform better with a 6 percent difference for 12-lead ECG and a 12 percent difference for CineECG.

We use the 3-class models in the further experiments with attribution and expert annotation due to 1. the better performance of the models on this task, which should lead to more accurate explanations, and 2. to reduce the number of tasks to be evaluated by the experts.

6.1.1. Model Attributions

Figure 3 presents the mean signals and attributions for each of the 3 classes. The strongest attributions appear in the V1, V3, II, and aVR leads. The mean signal data in leads I, II, and aVR appear to strongly influence the model towards the *normal* class prediction. Attributions for the *left side conduction disorders* are the strongest in the V1 (highest overall attribution score), V3, V5, and V6 leads, while V1, V3, and V4 have the strongest evidence for the *right side CDs*.

The attributions for all of the classes are mainly dispersed throughout the 12 leads. This result suggests that the various projections of the heart signal can capture different characteristics of the heart

cycle, thus enhancing the performance of the model trained on that data. The V1 lead is the only one possessing strong attributions towards all three of the classes; they are not well separated, particularly in the case of the *normal* and *left CD* classes, for which the attributions are practically located within the same timestep ranges.

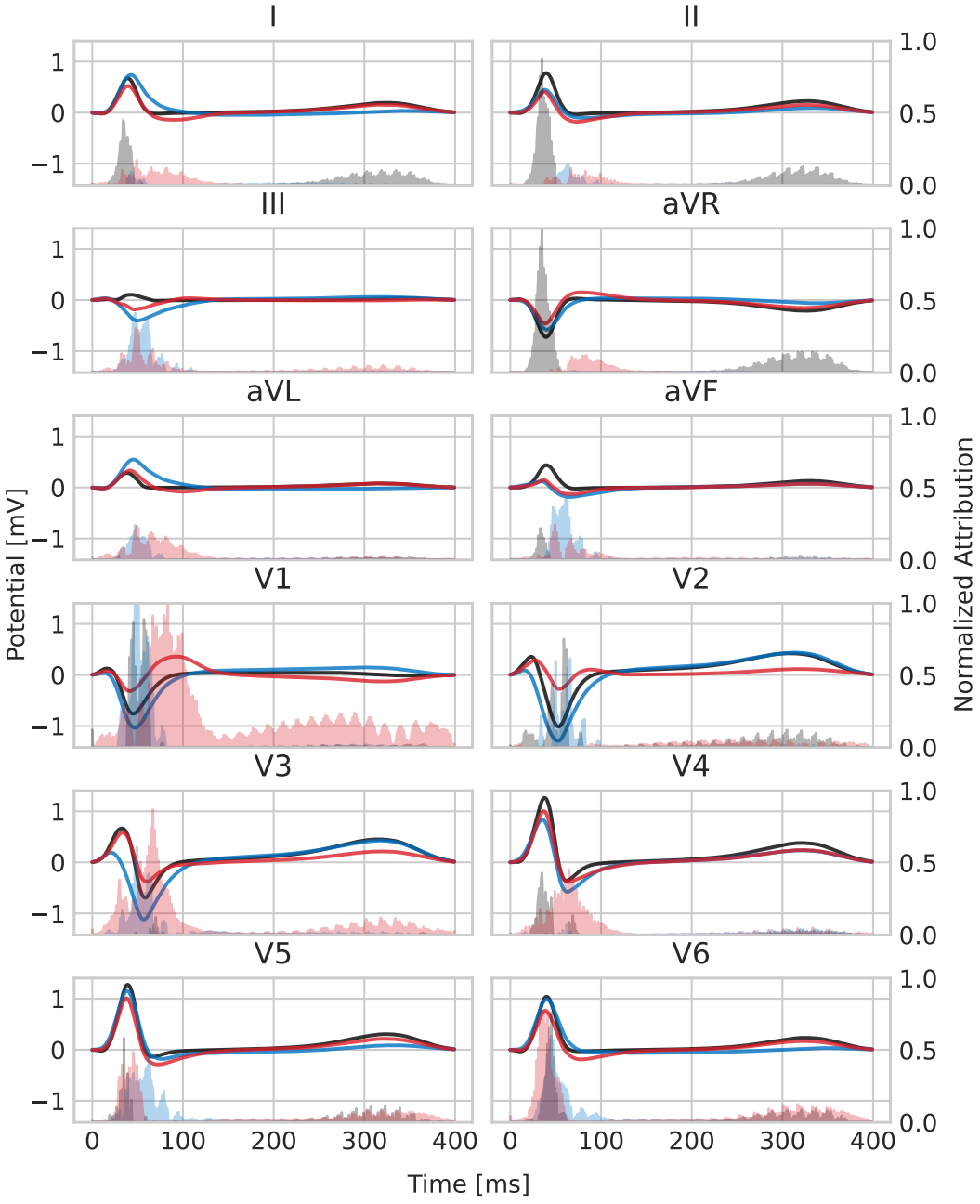


Figure 3: Mean signals and integrated gradients attributions for the 3-class 12-lead ECG model. Black – normal, blue – left CDs, and red – right CDs.

The attributions for the CineECG model presented in Figure 4 are less varied due to the lower number of dimensions. The X and Z axes of the CineECG ventricular representation show attributions that are dispersed throughout the timestep range. The X axis shows strong attributions toward the *left CD* class with multiple peaks in the beginning and the end of the signal, as well as slight evidence for the *right CD* class throughout the signal. The attributions in the Z axis are similar to those in the X axis, although considerably smaller in magnitude.

The Y axis of the CineECG representation shows the most diverse attributions: The *neutral* class

has a strong peak in attributions around the 50 ms; the left CD class has two peaks, around 40 and 60 milliseconds; and the right CD class has a strong attribution with a peak at approximately 75th millisecond. This result shows that the Y axis in the CineECG data representation is likely to be the most important feature for the classification of those disorders. This finding is in line with the clinical importance of this axis in the diagnosis of conduction disorders, as it corresponds to the left-right axis of the human heart.

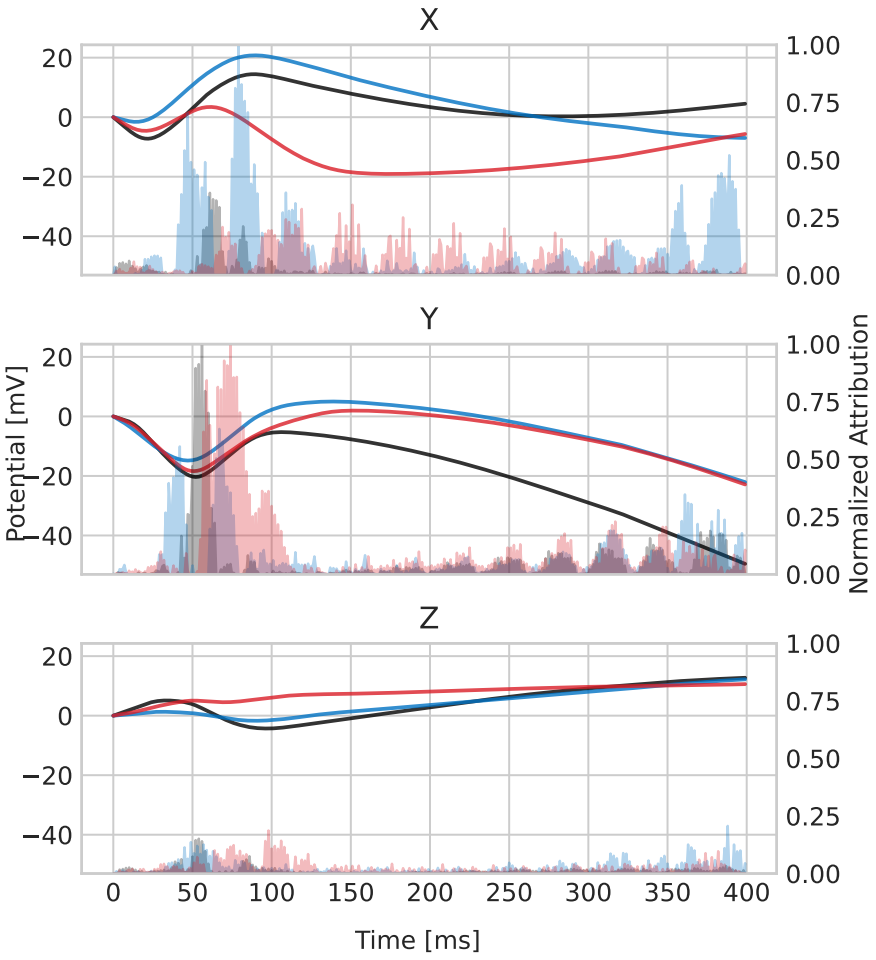


Figure 4: Mean trajectories and integrated gradients attribution for the 3-class CineECG model. Black – normal, blue – left CDs, and red – right CDs.

6.1.2. Ablation: CineECG Axes

To further test the finding concerning the importance of the axes in the CineECG ventricular representation, we conduct an ablation study by comparing the performance of a model trained on each axis independently to the performance of the original CineECG model.

We train and evaluate the models along the same lines as for the models in the previous Section. We obtain the following results for the three models: X-axis with F1-score and accuracy of 0.64; Y-axis with F1-score and accuracy of 0.75; and Z-axis with F1-score of 0.64 and accuracy of 0.65.

This result confirms the previous finding in the attribution analysis and the verdict that the Y-axis of the CineECG carries the most, if not all, of the predictive power of the data representation in the classification of left/right conduction disorders.

6.2. Expert Preferences

We present the results of the expert evaluation in Figure 5 as mean metric scores per explanation type. Overall, the median score was 4 for all three qualitative metrics. This score was also the most common response, accounting for 53.1%, 56.2%, and 71.9% of the ratings for clinical agreement, informativeness, and interpretability, respectively. This indicates that, for the most part, experts found the explanations clinically plausible, useful, and easy to interpret.

The distribution of modality reliance is shown separately in Figure 6. Modality reliance was skewed toward the conventional 12-lead ECG representation: 40.6% of the evaluations relied exclusively on 12-lead ECG, and 31.2% relied mostly on 12-lead ECG. In contrast, 18.8% of the evaluations were balanced between both modalities, 9.4% relied mostly on CineECG, and none relied exclusively on CineECG.

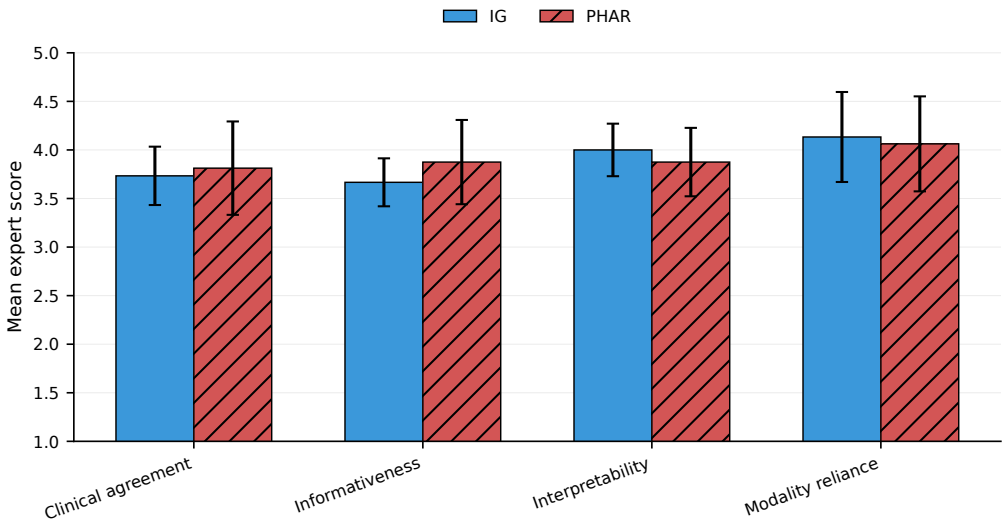


Figure 5: Mean expert scores by explanation type. Bars represent the 95% confidence intervals.

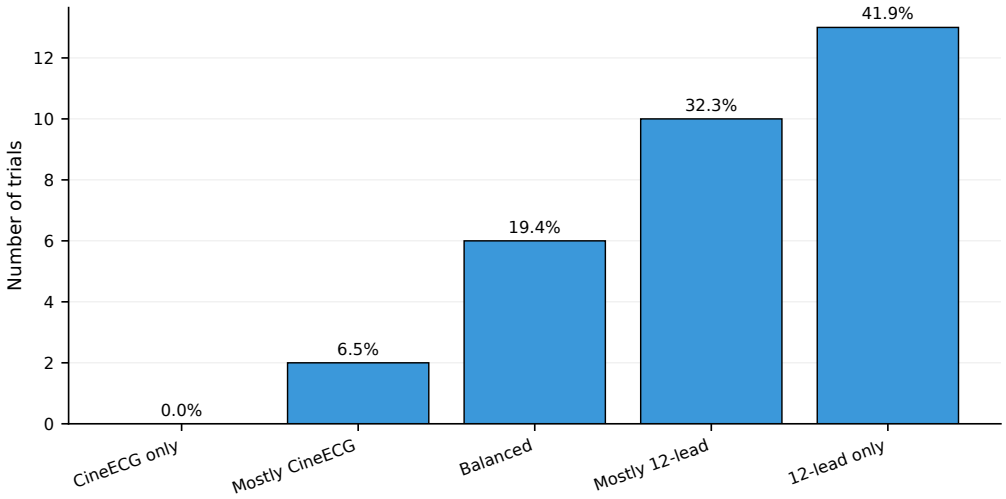


Figure 6: Distribution of expert reliance on data representation across 32 trial-level evaluations.

The mean differences between PHAR and IG were small. PHAR received slightly higher scores for clinical agreement (3.81 vs. 3.75) and informativeness (3.88 vs. 3.69), whereas IG received a slightly higher score for interpretability (4.00 vs. 3.88). Modality reliance was nearly identical for both explanation types (4.06 for PHAR and 4.00 for IG), indicating that the explanation type did not substantially

affect reliance on 12-lead ECG versus CineECG.

Explanation	Class	Clin. agr.	Inform.	Interp.	Mod. rel.
Combined	Normal	3.62	3.62	4.00	3.75
	Left CD	3.92	3.67	3.92	4.17
	Right CD	3.75	4.00	3.92	4.08
PHAR	Normal	4.00	4.00	3.67	2.67
	Left CD	4.00	3.50	3.83	4.17
	Right CD	3.57	4.14	4.00	4.57
IG	Normal	3.40	3.40	4.20	4.40
	Left CD	3.83	3.83	4.00	4.17
	Right CD	4.00	3.80	3.80	3.40

Table 4
Mean annotation scores aggregated by explanation type and ground-truth class. Modality reliance is encoded from 1 (*CineECG only*) to 5 (*12-lead only*).

Analyzing the results per ground-truth class (Table 4, "Combined"), we observe broadly similar clinical agreement and interpretability scores across the three classes. Left-CD cases received the highest mean clinical agreement score, whereas right-CD cases received the highest informativeness score. The most noticeable deviation in modality reliance appears for normal cases under PHAR, where the mean score of 2.67 lies between *mostly CineECG* and *balanced*. This result should be interpreted with caution, as this cell contains only three trials; however, it suggests that CineECG can serve as a complementary approach when the standard 12-lead ECG lacks defining features of a conduction disorder, or as an adjunctive modality to further reinforce the diagnosis.

Free-text comments further support this interpretation. As shown in Figure 7, experts most frequently referred to QRS morphology and duration, followed by lead-specific evidence from conventional ECG leads, particularly V1, V2, and V6. Mentions of conduction-disorder terminology, axis deviation, and notching or slurring indicate that the evaluations were grounded in standard diagnostic ECG reasoning. The relatively rare references to CineECG or the vector-based interpretation suggest that CineECG was used mainly as a complementary representation rather than as the primary basis for expert judgment.

Table 5 shows descriptive Pearson correlations between the annotation metrics. Informativeness has significant positive correlations with clinical agreement ($r = 0.43$, $p = 0.0139$) and interpretability ($r = 0.56$, $p = 0.0009$), whereas the correlation between clinical agreement and interpretability is weak. This is not unexpected, as an explanation may be easy to interpret without necessarily agreeing with clinical reasoning, and vice versa.

Modality reliance is not strongly correlated with any of the other three metrics, giving no evidence that either data representation yields systematically better explanations.

	Clin. agr.	Inform.	Interp.	Mod. rel.
Clin. agr.	1	0.43^{*}	0.10	-0.15
Inform.	0.43^{*}	1	0.56^{***}	-0.04
Interp.	0.10	0.56^{***}	1	0.16
Mod. rel.	-0.15	-0.04	0.16	1

Table 5
Pearson correlations between annotation metrics computed over 32 trial-level evaluations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

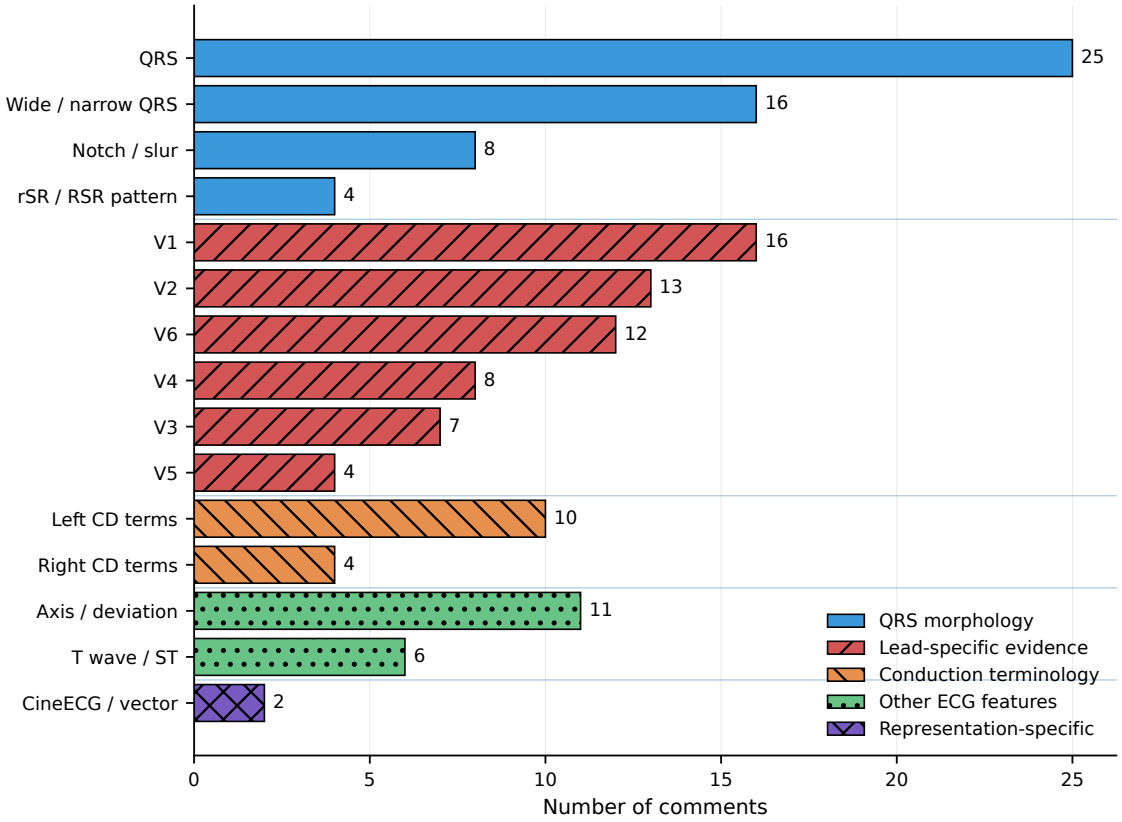


Figure 7: Clinical concepts mentioned in expert free-text comments, grouped by concept type. Counts indicate the number of comments in which each concept category was identified.

7. Discussion

Overall, our results indicate that the 12-lead ECG remained the dominant representation for both model performance and expert reasoning. Models trained on 12-lead ECG outperformed CineECG models in both classification settings. This suggests that the full 12-lead signal preserves more task-relevant morphology than the CineECG ventricular trajectory mapping. The reduced 3-class task provided a more stable basis for the explanation and expert-evaluation experiments because it achieved higher model performance and preserved the clinically relevant distinction between normal, left-sided, and right-sided conduction-disorder cases while reducing the data annotation task.

The results of our experiments indicate that CineECG retains clinically meaningful information, but in a more compressed and less redundant form than the full 12-lead ECG representation. The attribution analysis and axis-wise ablation showed that the CineECG Y-axis carried most of the predictive signal. The expert evaluation also supports this interpretation. Experts primarily relied on the standard ECG view even when CineECG was available, and their free-text comments were mostly grounded in QRS morphology, lead-specific evidence, and standard conduction-disorder criteria. Thus, CineECG appears more useful as a complementary view than as a replacement for conventional ECG.

Comparing the clinical utility of PHAR validity corridors and Integrated Gradients heatmaps showed only small differences. PHAR was rated slightly higher for clinical agreement and informativeness, while Integrated Gradients was rated slightly higher for interpretability. This suggests that the two explanation types support different aspects of expert reasoning. Integrated Gradients provides simple localization of salient features that help experts see where the model assigns importance, while on the other hand, PHAR provides explicit validity corridors that can be inspected against the signal values and interpreted as rule-like constraints. Clinical XAI interfaces may therefore benefit from combining such complementary explanation forms rather than selecting one method as universally preferable.

The main limitations are the small expert cohort, the descriptive nature of the statistical analysis, and the fact that the experts were presented with both 12-lead ECG and CineECG in one view. This design allowed us to measure modality reliance, but it does not isolate the independent effect of each representation.

Future work should evaluate larger expert cohorts, separate modality conditions, additional model architectures, and more realistic clinical workflows in which diagnostic accuracy, decision confidence, and review time can be measured directly.

8. Conclusions

In this work, we compared the utility of two data representation types for conduction disorder classification: conventional 12-lead ECG and CineECG, and evaluated two explanation types: Integrated Gradients attributions and PHAR validity corridors. The results show that 12-lead ECG remains the stronger representation for both model performance and expert reasoning. CineECG did not replace the standard ECG view, but the attribution and ablation analyses indicate that it retains clinically meaningful information, especially along the Y-axis, and may therefore serve as a complementary representation. The expert evaluation showed positive but similar ratings for both explanation types: Integrated Gradients provided intuitive localization cues, while PHAR offered more explicit signal constraints that can be inspected against the waveform. These findings suggest that clinical XAI systems should not only introduce new representations or explanation methods, but also integrate them with established diagnostic views and concepts familiar to clinicians. Future work should evaluate these findings with larger expert cohorts, separate modality conditions, and more realistic clinical workflows.

Declaration on Generative AI

During the preparation of this work, the authors used Google Gemini Pro in: Grammar and spelling checking, proofreading, and coding assistance for computational tasks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Funding

This paper is part of a project that has received funding from the European Union's Horizon Europe Research and Innovation Programme, under Grant Agreement number 101120406. The paper reflects only the authors' views, and the EC is not responsible for any use that may be made of the information it contains.

The contribution of Karol Dobiczek for the research for this publication has been supported by a grant from the Priority Research Area (DigiWorld) under the Mark Kac Center for Complex Systems Research Strategic Programme Excellence Initiative at Jagiellonian University.

Acknowledgments

The authors would like to thank the expert evaluators Tomasz Więcek, Anna Kulczycka, and Adam Palarski for their valuable feedback and participation in the experiments.

The authors thank the Faculty of Physics, Astronomy and Applied Computer Science at Jagiellonian University for computational resources.

References

[1] J. Suh, J. Kim, S. Kwon, E. Jung, H.-J. Ahn, K.-Y. Lee, E.-K. Choi, W. Rhee, Visual interpretation of deep learning model in ECG classification: A comprehensive evaluation of feature attribution

methods, Computers in Biology and Medicine 182 (2024) 109088. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482524011739. doi:10.1016/j.compbiomed.2024.109088.

[2] A. Taleban, R. Sparapani, P. Noffke, S. Zlochiver, Q. Lu, M. E. Widlansky, J. Luo, Explainable artificial intelligence in electrocardiography: A systematic review, Biomedical Signal Processing and Control 114 (2026) 109325. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809425018361. doi:10.1016/j.bspc.2025.109325.

[3] B. K. O. Arends, W. A. C. Van Amsterdam, P. Van Der Harst, M. Van Smeden, R. Van Es, R. R. Van De Leur, Signal or noise? Evaluating commonly used attribution methods for explaining deep neural networks in electrocardiogram classification, European Heart Journal - Digital Health 7 (2026) ztag038. URL: https://academic.oup.com/ehjdh/article/doi/10.1093/ehjdh/ztag038/8512917. doi:10.1093/ehjdh/ztag038.

[4] A. Hudon, T. Demazure, A. Karran, P.-M. Léger, S. Sénécal, Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence, in: F. D. Davis, R. Riedl, J. Vom Brocke, P.-M. Léger, A. B. Randolph, G. Müller-Putz (Eds.), Information Systems and Neuroscience, volume 52, Springer International Publishing, Cham, 2021, pp. 237–246. URL: https://link.springer.com/10.1007/978-3-030-88900-5_27. doi:10.1007/978-3-030-88900-5_27, series Title: Lecture Notes in Information Systems and Organisation.

[5] M. J. Boonstra, B. N. Hilderink, E. T. Locati, F. W. Asselbergs, P. Loh, P. M. Van Dam, Novel *CineECG* enables anatomical 3D localization and classification of bundle branch blocks, EP Europace 23 (2021) i80–i87. URL: https://academic.oup.com/europace/article/23/Supplement_1/i80/6158557. doi:10.1093/europace/euaa396.

[6] S. Katsushika, S. Kodera, S. Sawano, H. Shinohara, N. Setoguchi, K. Tanabe, Y. Higashikuni, N. Takeda, K. Fujiu, M. Daimon, H. Akazawa, H. Morita, I. Komuro, An explainable artificial intelligence-enabled electrocardiogram analysis model for the classification of reduced left ventricular function, European Heart Journal - Digital Health 4 (2023) 254–264. URL: https://academic.oup.com/ehjdh/article/4/3/254/7125058. doi:10.1093/ehjdh/ztad027.

[7] O. Kovalchuk, O. Barmak, P. Radiuk, L. Klymenko, I. Krak, Towards Transparent AI in Medicine: ECG-Based Arrhythmia Detection with Explainable Deep Learning, Technologies 13 (2025) 34. URL: https://www.mdpi.com/2227-7080/13/1/34. doi:10.3390/technologies13010034.

[8] J. Ojha, H. Haugerud, A. Yazidi, P. G. Lind, Exploring Interpretable AI Methods for ECG Data Classification, in: The Fifth Workshop on Intelligent Cross-Data Analysis and Retrieval, ACM, Phuket Thailand, 2024, pp. 11–18. URL: https://dl.acm.org/doi/10.1145/3643488.3660294. doi:10.1145/3643488.3660294.

[9] M. Ganeshkumar, R. Vinayakumar, V. Sowmya, E. Gopalakrishnan, K. Soman, Explainable Deep Learning-Based Approach for Multilabel Classification of Electrocardiogram, IEEE Transactions on Engineering Management 70 (2023) 2787–2799. URL: https://ieeexplore.ieee.org/document/9537612/. doi:10.1109/TEM.2021.3104751.

[10] Y.-Y. Jo, Y. Cho, S. Y. Lee, J.-m. Kwon, K.-H. Kim, K.-H. Jeon, S. Cho, J. Park, B.-H. Oh, Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram, International Journal of Cardiology 328 (2021) 104–110. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167527320342224. doi:10.1016/j.ijcard.2020.11.053.

[11] H. Turbé, M. Bjelogrlic, C. Lovis, G. Mengaldo, Evaluation of post-hoc interpretability methods in time-series classification, Nature Machine Intelligence 5 (2023) 250–260. URL: https://www.nature.com/articles/s42256-023-00620-w. doi:10.1038/s42256-023-00620-w.

[12] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards a Rigorous Evaluation of XAI Methods on Time Series, 2019. URL: http://arxiv.org/abs/1909.07082. doi:10.48550/arXiv.1909.07082, arXiv:1909.07082 [cs].

[13] D. Mercier, J. Bhatt, A. Dengel, S. Ahmed, Time to Focus: A Comprehensive Benchmark Using Time Series Attribution Methods, 2022. URL: http://arxiv.org/abs/2202.03759. doi:10.48550/arXiv.2202.03759, arXiv:2202.03759 [cs].

[14] K. Dobiczek, M. Mozolewski, S. Bobek, M. Szafarczyk, P. v. Dam, G. J. Nalepa, Validating the Clinical Utility of CineECG 3D Reconstructions through Cross-Modal Feature Attribution, 2026.

URL: http://arxiv.org/abs/2604.27017. doi:10.48550/arXiv.2604.27017, arXiv:2604.27017 [eess].

[15] A. B. Haque, A. N. Islam, P. Mikalef, Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research, Technological Forecasting and Social Change 186 (2023) 122120. URL: https://linkinghub.elsevier.com/retrieve/pii/S0040162522006412. doi:10.1016/j.techfore.2022.122120.

[16] M. Mozolewski, S. Bobek, G. J. Nalepa, Explaining time series classifiers with PHAR: Rule extraction and fusion from post-hoc attributions, 2025. URL: https://arxiv.org/abs/2508.01687. doi:10.48550/arXiv.2508.01687. arXiv:2508.01687, earlier preprint version.

[17] M. Mozolewski, B. Bayrak, K. Bach, G. J. Nalepa, From prototypes to sparse ECG explanations: SHAP-driven counterfactuals for multivariate time-series multi-class classification, Information Systems Frontiers (2026). URL: https://doi.org/10.1007/s10796-026-10711-9. doi:10.1007/s10796-026-10711-9, special Issue: Explainability in Human-Centric AI.

[18] M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks, 2017. URL: http://arxiv.org/abs/1703.01365. doi:10.48550/arXiv.1703.01365, arXiv:1703.01365 [cs].

[19] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocardiography dataset, 2022. URL: https://physionet.org/content/ptb-xl/1.0.3/. doi:10.13026/KFZX-AW45.