

From Prototypes to Sparse ECG Explanations: SHAP-Driven Counterfactuals for Multivariate Time-Series Multi-class Classification

Maciej Mozolewski^{1,3*†}, Betül Bayrak^{2†}, Kerstin Bach²,
Grzegorz J. Nalepa^{3,4}

¹ Jagiellonian Human-Centered AI Lab, Mark Kac Center for Complex Systems Research, Jagiellonian University, Łojasiewicza 11, Kraków, 30-348, Poland .

²Department of Computer Science, Norwegian University of Science and Technology (NTNU), Høgskoleringen 1, Trondheim, 7034, Norway.

³Department of Human-Centered Artificial Intelligence, Institute of Applied Computer Science, Jagiellonian University, Łojasiewicza 11, Kraków, 30-348, Poland .

⁴School of Information Technology, Halmstad University, Halmstad, Sweden.

*Corresponding author(s). E-mail(s): m.mozolewski@uj.edu.pl;
Contributing authors: betul.bayrak@ntnu.no; kerstin.bach@ntnu.no;
grzegorz.j.nalepa@uj.edu.pl;

[†]These authors contributed equally to this work.

Abstract

In eXplainable Artificial Intelligence (XAI), instance-based explanations for time series have gained increasing attention due to their potential for actionable and interpretable insights in domains such as healthcare. Addressing the challenges of explainability of state-of-the-art models, we propose a prototype-driven framework for generating sparse counterfactual explanations tailored to 12-lead ECG classification models. Our method employs SHAP-based thresholds to identify critical signal segments and convert them into interval rules, uses Dynamic Time Warping (DTW) and medoid clustering to extract representative prototypes, and aligns these prototypes to query R-peaks for coherence with the sample being explained. The framework generates counterfactuals that modify only 78% of the original signal while maintaining 81.3% validity across all classes and

achieving 43% improvement in temporal stability. We evaluate three variants of our approach, Original, Sparse, and Aligned Sparse, with class-specific performance ranging from 98.9% validity for myocardial infarction (MI) to challenges with hypertrophy (HYP) detection (13.2%). This approach supports near real-time generation (< 1 second) of clinically valid counterfactuals and provides a foundation for interactive explanation platforms. Our findings establish design principles for physiologically-aware counterfactual explanations in AI-based diagnosis systems and outline pathways toward user-controlled explanation interfaces for clinical deployment.

Keywords: Explainable AI, ECG explanations, Sparse time-series counterfactuals, Rule extraction

1 Introduction

One of the key challenges in healthcare engineering is the automatic analysis of ECG signals for early detection of cardiac abnormalities and timely clinical intervention. Despite digital recording, interpretation of 12-lead ECGs still depends on expert visual assessment of waveforms and remains challenging due to inter-individual variability in gender, electrode placement, and electrophysiological properties (van Dam, Boonstra, Locati, & Loh, 2021). Conversely, recent innovations in healthcare have made the ECG increasingly a commodity, while training in ECG interpretation remains limited, placing a greater workload on experts and potentially hampering the early detection of cardiac patients (Proniewska, Abächerli, & van Dam, 2023).

On the other hand, the availability of large-scale, clinically accurate 12-lead ECG datasets (e.g., PTB-XL (Wagner, Strodthoff, Bousseljot, Samek, & Schaeffter, 2022; Wagner et al., 2020) and PTB-XL+ (Strodthoff et al., 2023)) has enabled the development of high-performance deep learning classifiers capable of detecting multiple cardiac conditions within a single recording. However, the inherent complexity of these models renders their decision-making processes opaque to end users (Holzinger, Langs, Denk, Zatloukal, & Müller, 2019), underscoring the need for tailored explainable AI techniques in ECG interpretation.

Despite progress in SHAP-based local explanations such as WindowSHAP (Nayebi, Tipirneni, Reddy, Foreman, & Subbian, 2023), current methods might be too slow for real-time deployment (Van Looveren & Klaise, 2021). Counterfactual generation techniques frequently violate ECG morphology and temporal coherence, yielding physiologically implausible waveforms. Baig, Nasir, Khan, and Haque (2025) describe a lightweight ECG analysis system tailored for wearable and embedded platforms. They highlight the demand for efficient models capable of real-time analysis in resource-limited settings, yet many XAI techniques still introduce substantial computational overhead. This persistent shortfall in explainability performance motivates our work: our framework delivers potentially low-latency inference, actionable counterfactual explanations, and the potential for point-of-care deployment on standard clinical workstations.

In recent works, explainable AI methods have been applied to ECG interpretation with promising results. Image-based techniques, such as those of [Ao and He \(2023\)](#), overlay heatmaps on 12-lead ECG images to highlight diagnostically relevant waveform features and improve model interpretability. They further advocate for advanced XAI strategies: counterfactual explanations and rule-based methods, to boost both transparency and robustness. Similarly, [Tanyel et al. \(2025\)](#) employed DiCE-based counterfactuals [Mothilal, Sharma, and Tan \(2020\)](#) on the PTB-XL dataset, validating generated examples with two cardiologists and demonstrating their clinical relevance. Building on these complementary approaches, our framework combines rule-based explanations, heatmaps and efficient counterfactual generation to deliver real-time, actionable insights.

In the context of ECG counterfactual explanations, *sparsity* refers to the proportion of the signal that requires modification to achieve a class transition. A sparse counterfactual modifies only a minimal subset of time points across the 12 leads, rather than perturbing the entire recording. This sparsity is crucial for clinical interpretability since sparse explanations highlight specific waveform segments that drive diagnostic changes instead of presenting wholesale signal replacements. By focusing clinical attention on these critical regions, which states a challenge in the literature, sparse counterfactuals align with how cardiologists naturally interpret ECGs—examining specific morphological features rather than global waveform differences.

Our framework addresses these challenges by combining SHAP-derived rule extraction and prototype-based sparse counterfactual generation. First, we compute SHAP values and select features above a global threshold, forming concise interval rules. Next, we measure distances between rule-defined series using Dynamic Time Warping (DTW) ([Sakoe & Chiba, 1978](#)) and employ clustering followed by medoid selection to extract representative prototypes for each class. Finally, we apply our optimization step, which consists of aligning and scaling these prototypes to the query ECG by matching characteristic R-peaks. To enrich comprehensibility, we interviewed clinicians, including cardiology and electrophysiology experts, using semi-structured, one-on-one sessions. Participants reviewed lead-wise importance charts, signal overlays, and both single- and multi-overlay counterfactual examples annotated with interval rules. They rated readability, design clarity, and information density, and offered improvement suggestions.

In summary, our method delivers fast, physiologically valid counterfactual ECG explanations. It combines SHAP-derived interval rules with precomputed prototypes to preserve both waveform morphology and global rhythm. By precomputing prototypes, counterfactuals are generated in near real time with the cost of querying and optimization. Clinician interviews confirmed the readability, clarity, and appropriate information density of our visualizations. This approach strengthens transparency and trust in AI-driven ECG analysis.

The main contributions of this work are as follows:

- A novel counterfactual generation approach that proposes sparse counterfactuals from prototypes derived from existing data samples, with optimization applied.

- A prototype-based strategy enabling real-time application by using pre-computed instances combined with an optimization process, resulting in tailored explanations with reduced computational cost.
- A quantitative evaluation framework specialized on ECG counterfactual explanations that covers validity, sparsity, stability, and margin metrics.
- A visualization framework integrating SHAP explanations with single- and multiple-counterfactual overlays, qualitatively evaluated with three cardiology experts.

The remainder of this article is structured as follows. Section 2 reviews the background and related work. Section 3 describes our proposed method for generating counterfactual explanations of ECG signals, including adaptation of the *Post-hoc Attribution Rule extraction (PHAR)* framework (Section 3.1), prototype selection (Section 3.2), and the explanation generation process (Section 3.3). Section 4 presents the experimental evaluation, including quantitative results and qualitative interviews with ECG analysts. Section 5 discusses the advantages and limitations of our approach and outlines directions for future work. Finally, Section 6 concludes the article by summarizing the main contributions. Additional details are provided in the appendices: Appendix A describes the details of ECG model training and calibration, Appendix B reports the metrics for ECG counterfactual evaluation, and Appendix C presents example ECG counterfactual visualizations from the proposed method.

2 Background

2.1 XAI in Healthcare

Recent work by Röber, Goedhart, and Birbil (2025) emphasizes the growing adoption of AI and XAI systems in healthcare. In the study’s semi-structured interviews, clinicians identified feature-importance scores as among the most immediately interpretable explanation methods. These visualizations, often simple bar charts, enable practitioners to quickly recognize which patient characteristics most influence the model’s prediction without requiring understanding of complex algorithmic techniques such as SHAP or LIME. However, clinicians also cautioned that this simplicity can be misleading. Some remarked that interpreting these scores may feel deceptively straightforward, raising concerns about oversimplification. To establish trust, they emphasized the need for additional contextual information, such as uncertainty estimates or reference distributions, to support these explanations (Röber, Lumadjeng, Akyüz, & Birbil, 2025).

While counterfactual and rule-based explanations offer deeper contextual understanding, they demand additional time and cognitive effort to assess. Visual representations, such as bar charts and signal overlays, are essential for translating model outputs into comprehensible insights at the point of care. Successful deployment depends on solutions that integrate seamlessly into existing workflows and on multidisciplinary collaboration between developers, researchers, and healthcare professionals.

Holzinger et al. (Holzinger, Langs, Denk, Zatloukal, & Müller, 2019) argue that effective XAI in medicine requires intuitive explanation interfaces and tight human-in-the-loop integration to ensure clinical usability. They emphasize the trade-off between ante-hoc models, which offer built-in transparency but may sacrifice accuracy, and post-hoc methods that provide flexible explanations for black-box predictors. Moreover, they advocate for causal frameworks capable of generating counterfactual insights, thereby aligning technical explainability with clinicians’ causal reasoning.

Salih et al. (2024) review 213 studies on XAI in cardiology and report that 43% of works do not validate explanations, 37% rely solely on literature comparison, 11% involve clinical experts, and 11% apply statistical or proxy-grounded evaluation. Commonly used techniques for ECG interpretation include SHAP (global and local importance scores) and Grad-CAM (activation maps), sometimes augmented with LIME. In the minority of studies engaging cardiologists, evaluation focuses on the alignment between model-highlighted critical points and expert interpretation. The authors emphasize the necessity of application-grounded validation with cardiologists, the adoption of proxy-grounded methods (which assess explanation fidelity by removing attributed features and retraining the model), permutation tests (which quantify the effect of feature shuffling on predictions) and a hybrid evaluation strategy combining human-, proxy-, and literature-grounded approaches to ensure clinical relevance and trustworthiness. A similar comprehensive evaluation approach is adopted by Darias, Díaz-Agudo, and Recio-Garcia (2025) and Darias, Bayrak, Caro-Martínez, Díaz-Agudo, and Recio-Garcia (2024), who employ addition and deletion metrics that systematically assess explanation fidelity by removing or adding attributed features and measuring the resulting impact on model predictions. These proxy-grounded methods provide quantitative measures of explanation quality by evaluating how feature modifications affect classification performance, complementing human-centered evaluation approaches. Such multi-faceted evaluation frameworks are essential for establishing the reliability and clinical utility of XAI systems in healthcare applications.

2.2 Interactive Visualization for Expert Feedback

Interactive machine learning (iML) highlights the critical role of end-user engagement through iterative model refinement. Wondimu et al. (Wondimu, Buche, & Visser, 2022) demonstrate that interactive visualizations, such as multi-view decision path diagrams and activation heatmaps, empower users to inspect model behavior, adjust parameters in real time, and directly observe the impact of their inputs. The authors emphasize that transparency, user control, and ongoing human-machine collaboration are essential foundations for trustworthy AI in high-stakes domains.

Theissler et al. (Theissler, Spinnato, Schlegel, & Guidotti, 2022) propose a four-fold taxonomy of XAI for time-series: time-point attribution, subsequence methods, instance-based strategies (including counterfactuals), and rule/logic frameworks. They observe that ECG applications predominantly rely on attribution heatmaps and subsequent highlighting, yet often lack interactive evaluation and standardized benchmarks. The authors emphasize visualization (saliency maps, activation heatmaps, annotated overlays) as essential for translating model decisions into experts’ insights.

Key research gaps include the development of higher-order and truly model-agnostic explanations, domain-specific interpretability techniques, user-friendly interfaces, and advanced evaluation protocols.

2.3 Deep Learning for ECG Classification

Deep neural networks achieve high accuracy on multivariate time series tasks (F. Wang et al., 2025). Classification models have shown state-of-the-art performance. An exemplary deep learning model for ECG analysis is described by Hannun et al. (2019), which achieves cardiologist-level arrhythmia detection but does not incorporate any explainability (XAI) techniques. Baig et al. (2025) develop lightweight 1D CNN models for on-device arrhythmia classification and integrate popular XAI methods: Grad-CAM to produce local activation maps that highlight key signal components (e.g., QRS complexes, T waves) and SHAP to assign per-timepoint attribution scores across different diagnoses. While these well-established techniques yield valuable model insights, their raw visualizations can be cluttered and insufficiently streamlined for fast clinical interpretation.

2.4 Counterfactual Explanations

Delaney, Greene, and Keane (2021) introduces Native Guide, an instance-based method for obtaining counterfactual explanations in time-series classification. The approach retrieves the nearest opposite-class exemplar from a reference database and adapts its key subsequences using classifier-derived weight vectors, yielding counterfactuals that align closely with the original data distribution. This exemplar search can be computationally demanding. The authors validate their method on multiple UCR benchmarks, including ECG200¹ (Olszewski, Maxion, & Siewiorek, 2001), which contains pre-segmented single-beat ECG signals. These simplified datasets are less complex than real-world clinical recordings. They report superior performance over gradient-based counterfactual techniques in terms of proximity, sparsity, plausibility, and diversity.

Van Looveren and Klaise (2021) introduces a model-agnostic method for generating counterfactual explanations in which the search is accelerated and interpretability is enhanced by guiding optimization with class prototypes. In their approach, prototypes are derived either as latent-space centroids from an encoder or via class-specific k-d tree medoids, effectively replacing costly gradient-based searches. The method was demonstrated on standard benchmarks, such as the MNIST image dataset and the Adult Census tabular dataset.

3 From SHAP-derived Rules to Sparse ECG Explanations

Our proposed method integrates rule-based feature selection with prototype-guided counterfactual generation to provide interpretable and clinically plausible explanations for ECG classification. As illustrated in Figure 1, the overall pipeline consists of three

¹<https://www.timeseriesclassification.com/description.php?Dataset=ECG200>

main components: (i) extraction of concise interval rules from SHAP importance values with *Post-hoc Attribution Rule extraction* (Mozolewski, Bobek, & Nalepa, 2025), (ii) prototype selection via medoid clustering in DTW-based Multi-Dimensional Scaling (MDS) space, and (iii) an explanation module that aligns selected prototypes with the input signal and generates sparse counterfactuals. This modular structure enables fast and physiologically coherent explanations suitable for point-of-care use.

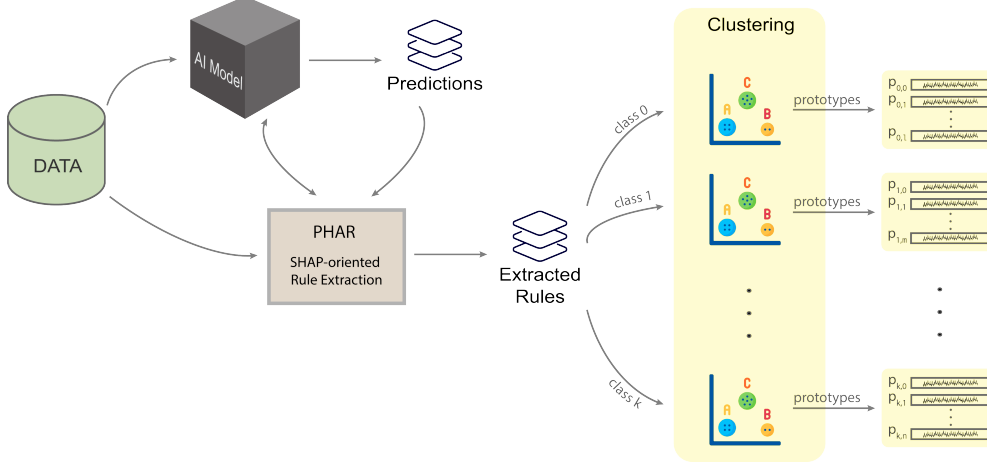


Fig. 1 Overview of the proposed framework.

3.1 Extracting Rules from SHAP Values

We use *Post-hoc Attribution Rule extraction (PHAR)*, a model-agnostic framework developed by Mozolewski et al. (2025) for converting numeric feature attributions into concise, instance-specific rules. *PHAR* supports numeric attributions from methods like SHAP (Lundberg & Lee, 2017) or LIME (Ribeiro, Singh, & Guestrin, 2016). First, we select top features via a global threshold of the absolute attributions at a chosen percentile. Then we apply controlled perturbations within each feature’s variance, identifying stable value intervals that preserve the original prediction. The generated rules may facilitate transparent decision-making and support domain experts in validating model behavior, and can be visualized on time series as semi-factual scenarios (Aryal & Keane, 2023). The full *PHAR* pipeline, including rule fusion and optimization, is detailed in the original publication and is omitted here for brevity.

In this work, however, the rules from *PHAR* serve to determine the most important features in ECG signals and are subsequently used for clustering. In Section 3.2, these rules are used to generate prototypes by clustering the rule sets for each class and selecting medoids, where the ECG instance corresponding to a medoid rule becomes the prototype and serves as a counterfactual.

3.1.1 Conversion of Numeric Importance Values to Rules

This procedure, adapted from the *PHAR* framework originally, concerns only the transformation of numeric feature attributions into instance-specific rules.

We have a set of N time series instances $\{X_n\}$ with model output vectors $\mathbf{y}_n$. We use each $\mathbf{y}_n$ as ground truth for explanations. For each instance n and feature f , the explainer (e.g., SHAP) gives a numeric importance $e_{n,f}$.

We collect all absolute importance values into the multiset

$$E = \{|e_{n,f}| : n = 1, \dots, N, f = 1, \dots, F\}. \quad (1)$$

We fix the global threshold T at the 90.00th percentile of E :

$$T = \text{Percentile}(E, 90.00). \quad (2)$$

A feature f is *important* for instance n if

$$|e_{n,f}| \geq T. \quad (3)$$

Let F_* be the set of important features for instance n .

For each $f \in F_*$, let $x_{n,f}$ be the original feature value. We compute the standard deviation σ_f of $\{x_{n,f} : n = 1, \dots, N\}$ on the training set.

We generate M perturbed samples jointly over all $f \in F_*$:

$$x_{n,f}^{(m)} \sim \text{Uniform}(x_{n,f} - \sigma_f, x_{n,f} + \sigma_f). \quad (4)$$

Each perturbed instance is fed to the model.

We then find, for each feature f , the smallest interval $(l_f, h_f]$ such that every perturbed sample produces the same full prediction vector $\mathbf{y}_n$. This enforces consistency across all output labels.

Finally, we define a rule for instance n :

$$R_n : \bigwedge_{f \in F_*} (x_f \in (l_f, h_f]). \quad (5)$$

3.2 Prototype Selection

In clinical ECG interpretation, practitioners require immediate explanations to support time-critical decisions. However, counterfactual generation methods that rely on exhaustive search or complex optimizations at inference time can introduce latency incompatible with point-of-care requirements. To address this challenge, we pre-compute representative prototypes for each diagnostic class to initialize counterfactual generation. Our prototype selection process consists of four key steps:

1. **Sample Filtering:** Let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$ be the training set where $x_i \in \mathbb{R}^{T \times C}$ is a 12-lead ECG signal and $y_i \in \{0, 1\}^L$ is its multi-label vector. For each class $l \in \{1, \dots, L\}$, we define:

$$\mathcal{D}_l = \{x_i : (x_i, y_i) \in \mathcal{D}_{\text{train}}, y_{i,l} = 1, \sum_{j=1}^L y_{i,j} = 1\} \quad (6)$$

This filters samples that belong exclusively to class l . This step excludes ambiguous or potentially misclassified recordings, yielding prototypes that are more stable, representative, and clinically meaningful.

2. **Distance-based Dimensionality Reduction:** In this step, we compute pairwise Dynamic Time Warping (DTW) distances between all training samples within each class, creating a distance matrix that captures morphological similarities between ECG signals.

For two ECG signals $x_i, x_j \in \mathcal{D}_l$, the Dynamic Time Warping distance is:

$$\text{DTW}(x_i, x_j) = \min_{\pi \in \Pi} \sum_{(t,s) \in \pi} \|x_i[t, :] - x_j[s, :]\|_2 \quad (7)$$

where Π is the set of all valid warping paths between the sequences.

For class l , we construct the distance matrix $D_l \in \mathbb{R}^{|\mathcal{D}_l| \times |\mathcal{D}_l|}$ where:

$$D_l[i, j] = \text{DTW}(x_i, x_j) \quad \forall x_i, x_j \in \mathcal{D}_l \quad (8)$$

We apply Multi-Dimensional Scaling to find a lower-dimensional embedding: $\min_{Z_l} \sum_{i,j} (\|z_i - z_j\|_2 - D_l[i, j])^2$ where $Z_l = [z_1, \dots, z_{|\mathcal{D}_l|}]^T \in \mathbb{R}^{|\mathcal{D}_l| \times d}$ and d is chosen to maximize:

$$s(d) = \frac{1}{|\mathcal{D}_l|} \sum_{i=1}^{|\mathcal{D}_l|} \frac{a(i) - b(i)}{\max(a(i), b(i))} \quad (9)$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance.

3. **Clustering in MDS Space:** We partition Z_l into k clusters using k-means:

$$\{C_1^l, \dots, C_k^l\} = \arg \min_{\mathcal{C}} \sum_{j=1}^k \sum_{z \in C_j} \|z - \mu_j\|_2^2 \quad (10)$$

For each cluster C_j^l , we select the medoid as the prototype, the prototype p_j^l for cluster j of class l is:

$$p_j^l = \arg \min_{x_i \in \mathcal{X}_{C_j^l}} \sum_{x_k \in \mathcal{X}_{C_j^l}} \text{DTW}(x_i, x_k) \quad (11)$$

where $\mathcal{X}_{C_j^l} = \{x_i : z_i \in C_j^l\}$ maps points back to original ECG space.

The final prototype set is $\mathcal{P} = \bigcup_{l=1}^L \bigcup_{j=1}^{k_l} \{p_j^l\}$.

3.3 From Prototypes to Sparse ECG Explanations

Given a query ECG signal and its model prediction, our method generates counterfactual explanations through a two-stage process: prototype alignment followed by sparsification. This approach ensures both temporal coherence and minimal perturbations.

3.3.1 R-peak Based Temporal Alignment

ECG signals exhibit natural variability in heart rate and rhythm, even within the same diagnostic class. To generate physiologically plausible counterfactuals that can be more easily analyzed by overlaying them, we align the temporal structure of selected prototypes to match the rhythm of the query signal. We employ R-peak detection and piecewise-linear warping:

1. **R-peak Detection:** R-peaks define the cardiac cycle boundaries and are detected in both the query signal and the selected prototype using either WFDB’s QRS algorithm or NeuroKit2’s peak detection, depending on signal quality.
2. **Peak Count Normalization:** If the prototype and query have different numbers of detected beats:
 - When the prototype has fewer beats, we prepend zeros to its start and, in the alignment step, only warp the beats that correspond to those in the query signal.
 - When the prototype has more beats, we trim it to match the query signal’s beat count.
3. **Piecewise-Linear Warping:** For each inter-peak segment and each ECG lead independently, we apply linear interpolation to map the prototype segment to the corresponding query segment. This preserves the morphology of individual cardiac cycles while adapting to the query’s specific rhythm.

The warping function for segment k between R-peaks is defined as:

$$\mathbf{x}_{\text{aligned}}[t_k : t_{k+1}] = \text{interp1d}(\mathbf{x}_{\text{proto}}[s_k : s_{k+1}], [0, 1] \rightarrow [0, 1]) \quad (12)$$

where t_k, t_{k+1} are consecutive R-peak positions in the query, and s_k, s_{k+1} are the corresponding peaks in the prototype.

3.3.2 Sparsity Optimization

The aligned prototype serves as an initial counterfactual, but it differs from the query at every time point. To enhance interpretability and identify the minimal changes required for class transition, we apply sparsity optimization:

1. **Importance scoring:** We compute the absolute difference between the aligned prototype and query at each time-channel point. R-peak regions receive doubled importance scores to preserve critical cardiac features.
2. **Iterative sparsification:** Starting with a low keep-ratio (e.g., 10%), we iteratively determine which signal segments to modify:

- Create a binary mask based on importance score percentiles
 - Replace only the masked regions of the query with the corresponding prototype values
 - Verify that the model’s prediction matches the target class
 - If successful, attempt a sparser solution; otherwise, increase the keep-ratio
3. **Segment Cleaning:** We enforce minimum segment lengths (10 samples) to avoid scattered, physiologically implausible modifications. Small isolated changes are removed from the modification mask.

The optimization objective is formulated as:

$$\min_{\mathbf{m}} \|\mathbf{m}\|_0 \quad \text{s.t.} \quad f(\mathbf{x}_q \odot (1 - \mathbf{m}) + \mathbf{x}_p \odot \mathbf{m}) = y_{\text{target}} \quad (13)$$

where $\mathbf{m}$ is the binary modification mask, $\mathbf{x}_q$ is the query, $\mathbf{x}_p$ is the aligned prototype, $\odot$ denotes element-wise multiplication, and f is the model.

3.3.3 Query Processing Workflow

The complete query processing workflow, illustrated in Figure 2, integrates the prototype selection methodology described in Section 3.2 with the alignment and sparsification techniques detailed above. This end-to-end pipeline transforms a query ECG and model prediction into an interpretable counterfactual explanation. When a new query ECG requires explanation:

1. Take the ECG signal, predicted classes by the AI model, and the target counterfactual class as input. If the target class is not provided, identify appropriate counterfactual classes based on the model’s prediction confidence. The system prioritizes transitions to neighboring classes in the diagnostic hierarchy (e.g., from pathological to normal) to keep clinical meaningfulness.
2. Retrieve the pre-computed prototypes for the target class, which were extracted during the offline MDS-based clustering phase. Select the prototype with minimal DTW distance to the query to choose a prototype that shares similar morphological characteristics with the input signal.
3. Apply the two-stage optimization process: (i) R-peak alignment to temporally warp the prototype, adapting its rhythm to match the query’s cardiac cycle, and (ii) sparsity optimization to identify the minimal set of modifications required for class transition while preserving physiological plausibility.
4. Return the sparse counterfactual along with visualizations highlighting the specific ECG regions that drive the diagnostic change, providing suitable visualizations to clinicians to focus on the most relevant waveform alterations.

This approach generates explanations in under a second by using pre-computed prototypes. The two-stage process, temporal alignment followed by sparse modifications, produces counterfactuals that respect ECG physiology. By preparing prototypes offline and running only lightweight optimizations at query time, we achieve both speed and quality, making the system practical for point-of-care use.

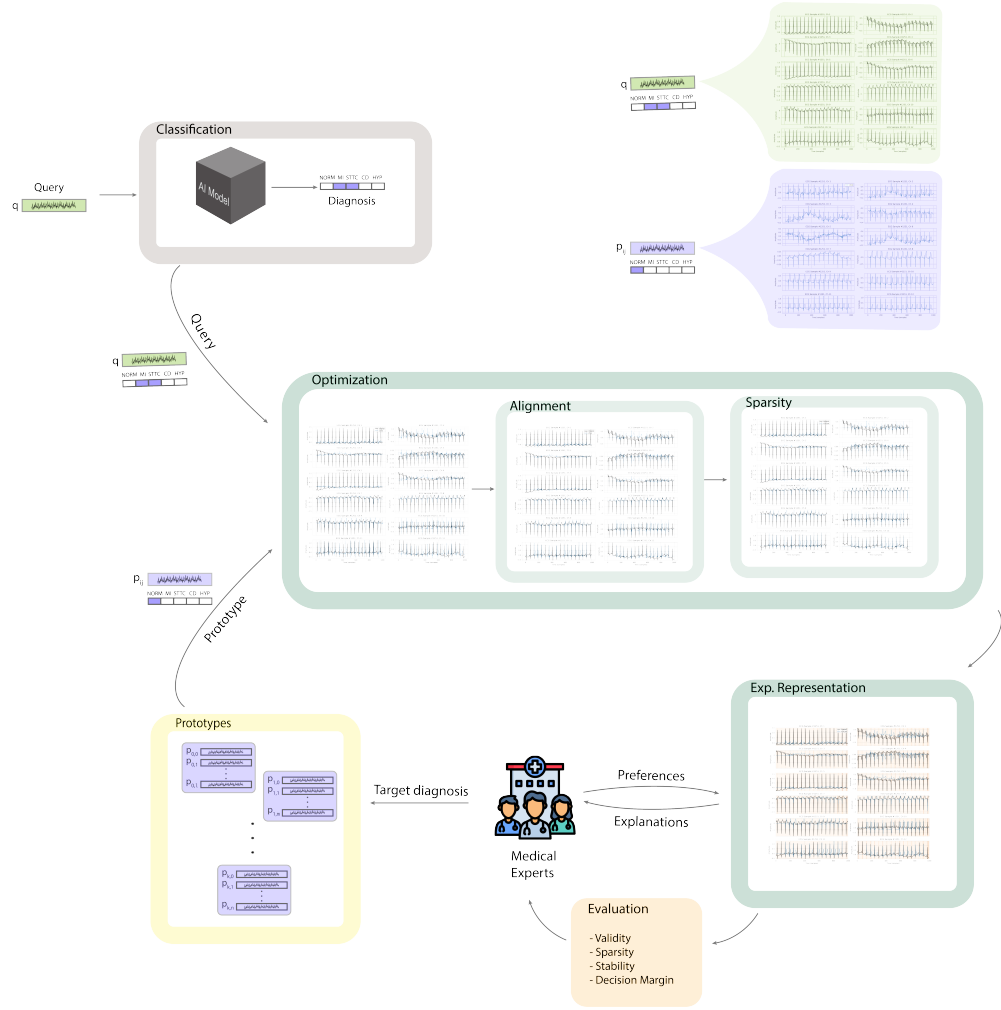


Fig. 2 ECG explanation generation pipeline: from query ECG classification through prototype retrieval, R-peak alignment, and sparsity optimization to final counterfactual visualization.

3.4 Objective Evaluation

To comprehensively assess the quality of generated counterfactual explanations, we developed five key evaluation metrics for 12-lead ECG explanations, which are commonly used in the literature for tabular counterfactuals (Bayrak and Bach (2024)), which we adapted for the specific characteristics of multivariate timeseries ECG signals. Unlike tabular data, where features are independent, ECG signals exhibit temporal dependencies, morphological constraints, and channel correlations that require specialized evaluation approaches.

3.4.1 Validity

Validity is the fundamental requirement for any counterfactual explanation, measuring whether it successfully changes the model’s prediction to the desired target class. For ECG signals, we must ensure that the counterfactual not only crosses the decision boundary but also maintains physiological plausibility. Given a query ECG signal $\mathbf{x} \in \mathbb{R}^{T \times C}$ (where $T = 1000$ samples and $C = 12$ leads) with model prediction $f(\mathbf{x}) = y$, and its counterfactual $\mathbf{x}'$, validity is defined as:

$$\text{validity}(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 & \text{if } f(\mathbf{x}') = y_{\text{target}} \wedge f(\mathbf{x}') \neq f(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

For multi-label ECG classification, we extend this to ensure all class predictions change appropriately:

$$\text{validity}_{\text{multi}} = \prod_{i=1}^L \mathbb{I}[f_i(\mathbf{x}') = y_{\text{target},i}] \quad (15)$$

where L is the number of diagnostic labels and f_i represents the i -th class prediction.

3.4.2 Sparsity

Sparsity quantifies the locality and efficiency of modifications, crucial for interpretability in clinical settings (Bayrak and Bach (2023)). Sparse counterfactuals highlight specific ECG features (e.g., ST-segment changes, QRS morphology) responsible for diagnostic decisions. We evaluate sparsity through multiple complementary metrics:

- **Sparsity Ratio** measures the fraction of modified samples across all channels:

$$\text{sparsity_ratio} = \frac{1}{T \times C} \sum_{t=1}^T \sum_{c=1}^C \mathbb{I}[|\mathbf{x}'_{t,c} - \mathbf{x}_{t,c}| > \tau] \quad (16)$$

where the threshold $\tau = 0.01 \cdot \sigma_{\mathbf{x}}$ accounts for noise tolerance, with $\sigma_{\mathbf{x}}$ being the standard deviation of the training data.

- **L_p Norms** provide alternative sparsity measures:

$$\text{sparsity}_{L_0} = \|\mathbf{x}' - \mathbf{x}\|_0 = \sum_{t,c} \mathbb{I}[|\mathbf{x}'_{t,c} - \mathbf{x}_{t,c}| > 0] \quad (17)$$

$$\text{sparsity}_{L_1} = \|\mathbf{x}' - \mathbf{x}\|_1 = \sum_{t,c} |\mathbf{x}'_{t,c} - \mathbf{x}_{t,c}| \quad (18)$$

$$\text{sparsity}_{L_2} = \|\mathbf{x}' - \mathbf{x}\|_2 = \sqrt{\sum_{t,c} (\mathbf{x}'_{t,c} - \mathbf{x}_{t,c})^2} \quad (19)$$

3.4.3 Stability

Stability metrics assess the robustness of counterfactual explanations against minor perturbations (Darias et al. (2024)), crucial for clinical reliability where measurement noise and physiological variations are inherent.

- **Noise Stability** tests robustness under Gaussian perturbations:

$$\text{stability_score} = \frac{1}{|\mathcal{N}| \cdot N_{\text{trials}}} \sum_{i=1}^{N_{\text{trials}}} \sum_{\sigma_{\epsilon} \in \mathcal{N}} \mathbb{I}[f(\mathbf{x}' + \epsilon_i) = f(\mathbf{x}')] \quad (20)$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon}^2 \mathbf{I})$ and $\mathcal{N} = \{0.01, 0.02, 0.05\} \cdot \text{std}(\mathbf{x}')$.

- **Temporal Stability** evaluates invariance to small time shifts, important for ECG signals where slight phase variations are common:

$$\text{temporal_stability} = \frac{1}{1 + \overline{\text{DTW}}_{\tau}} \quad (21)$$

where:

$$\overline{\text{DTW}}_{\tau} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \frac{\text{DTW}(\mathbf{x}', \text{shift}(\mathbf{x}', \tau))}{\text{DTW}_{\text{norm}}} \quad (22)$$

with $\mathcal{T} = \{-2, -1, 1, 2\}$ samples and $\text{DTW}_{\text{norm}} = \sqrt{T \cdot C}$ for normalization.

3.4.4 Decision Margin

The decision margin quantifies how confidently the counterfactual achieves the target classification, providing a buffer against classification boundary fluctuations:

$$\text{decision_margin} = p_{f(\mathbf{x}')}(\mathbf{x}') - t_{f(\mathbf{x}')} \quad (23)$$

where $p_{f(\mathbf{x}')}(\mathbf{x}')$ is the predicted probability and $t_{f(\mathbf{x}')}$ is the class-specific threshold (which may differ from 0.5 in multi-label settings).

These metrics collectively measure that the generated counterfactual explanations meet the requirements of clinical ECG interpretation, such as they must be valid, sparse enough to be interpretable, stable under realistic variations, and confident in their predictions. These evaluation metrics allow systematic comparison and selection of high-quality counterfactual explanations for 12-lead ECG classification models.

4 Experiments

4.1 Data

The PTB-XL dataset contains 21 799 clinical 12-lead ECG recordings sampled at 100 Hz as mentioned in Wagner et al. (2022). Each recording lasts 10s and includes patient metadata, signal-quality annotations, and diagnostic labels. PTB-XL is available via PhysioNet (Goldberger et al. (2000); Wagner et al. (2022)) and was first described in Scientific Data by Wagner et al. (2020).

Table 1 shows the distribution of the five main diagnostic classes. Normal ECGs (NORM) are most common, followed by myocardial infarction (MI), ST/T changes (STTC), conduction disturbances (CD) and hypertrophy (HYP).

Table 1 Distribution of primary diagnostic classes (n=21 799).

Class	Percentage (%)	Count
NORM	43.64	9 514
MI	25.09	5 469
STTC	24.01	5 235
CD	22.47	4 898
HYP	12.15	2 649

Abbreviations: NORM – Normal ECG; MI – Myocardial Infarction; STTC – ST/T Change; CD – Conduction Disturbance; HYP – Hypertrophy.

Table 2 lists the most frequent diagnostic label combinations of at least 2 labels in PTB-XL, which exceeds 1.0%. The most common pairing is MI with CD (8.23%), followed by STTC with HYP (6.92%) and MI with STTC (6.14%), which are most prevalent overlapping pathologies.

Table 2 Diagnostic label combinations exceeding 1 % in PTB-XL.

Combination	Count	Percentage (%)
MI, CD	1794	8.23
STTC, HYP	1509	6.92
MI, STTC	1339	6.14
STTC, CD	1066	4.89
MI, HYP	818	3.75
HYP, CD	787	3.61
MI, STTC, HYP	517	2.37
NORM, CD	415	1.90
MI, STTC, CD	379	1.74
STTC, HYP, CD	367	1.68
MI, HYP, CD	274	1.26

Abbreviations: NORM – Normal ECG; MI – Myocardial Infarction; STTC – ST/T Change; CD – Conduction Disturbance; HYP – Hypertrophy.

4.2 Data Preprocessing

The authors of PTB-XL provided stratified sample folds, which we adopted without modification. In our case, folds 1–7 was held out for training and fold 9 for validation. For fold 10 we calculated SHAP values, which then were used to create rules in *PHAR* method and further clustered to obtain medoids (which acted as counterfactuals). Additionally fold 7 was used for calculating statistics for *PHAR* method and fold 9 was used as a background for calculating SHAP values. Fold 8, not used in any other place, served for queries. This split yielded 15 245 training samples (69.9%), 2 183 validation samples (10.0%), 2 198 samples to create rules and clustering them (10.1%) and 2 173 query samples (10.0%).

All ECG signals were normalized using statistics computed on the training set. We found the mean of X_{train} to be $\mu = -0.0008313$ and its standard deviation $\sigma = 0.2357998$. Each signal x was then scaled as $x_{\text{norm}} = \frac{x - \mu}{\sigma + 10^{-7}}$.

Diagnostic labels were converted to multi-hot encodings for deep-network training. We applied a *multi-label binarizer* to map each list of labels into a fixed five-dimensional binary vector, where each position corresponds respectively to Normal ECG, Myocardial Infarction, ST/T Change, Conduction Disturbance and Hypertrophy. These multi-hot vectors served as the target outputs during model training.

4.3 Model

In Figure 3, we present the architecture of our 12-lead ECG multiclass classification model. The network accepts raw ECG inputs of shape $1000 \text{ time points} \times 12 \text{ leads}$. First, two convolutional blocks extract local waveform features; each block combines convolutional layers (*blue*), batch normalization (*purple*), ReLU activations (*pink*), and max-pooling (*green*). The encoded features are then fed into a GRU layer (*orange*) for temporal integration. Next, a RealNVP normalizing flow (*lavender*) transforms the GRU output into a latent space. Finally, an MLP head (*yellow*), consisting of two fully connected layers, maps the latent representation to class logits and produces probability scores via sigmoid activation.

4.3.1 Model architecture

The network comprises convolutional, recurrent, flow-based and feed-forward components. Specifically:

- **Input:** Raw ECG signal of shape (1000, 12).
- **Conv Block 1:** Conv1D with 16 filters (kernel size 5, padding 2), followed by batch normalization, ReLU activation and max-pooling, yielding output of shape (500, 16).
- **Conv Block 2:** Conv1D with 32 filters (kernel size 5, padding 2), followed by batch normalization, ReLU activation and max-pooling, yielding output of shape (250, 32).
- **GRU Layer:** Processes the sequence of length 250 with 32-dimensional inputs to produce a fixed embedding of size H .
- **RealNVP Flow:** Two coupling layers with alternating binary masks; each layer uses MLPs to scale and translate one half of the H -dimensional vector, producing a latent vector of the same size.

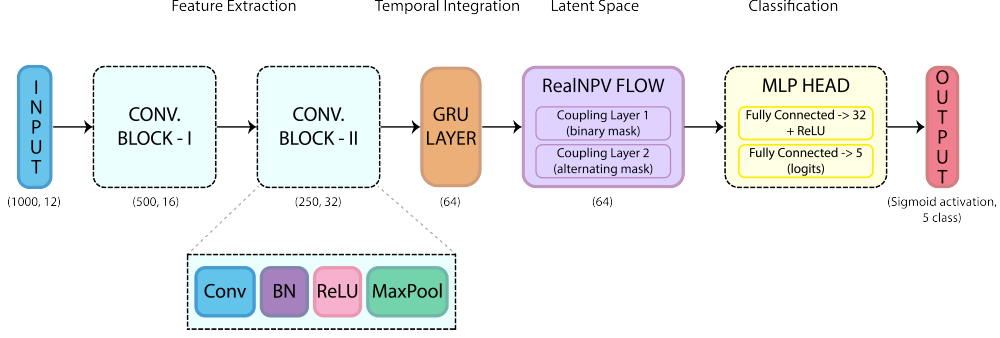


Fig. 3 12-lead ECG multiclass classification architecture: raw signals (1000 time points $\times$ 12 leads) are processed through convolutional feature extraction (blue), batch normalization (purple), ReLU activations (pink), and max-pooling (green); then temporally modeled by a GRU (orange), transformed in latent space via RealNVP flow (lavender), and classified with an MLP head (yellow).

- **MLP Head:** Fully connected layer to 32 units with ReLU, then a second layer to C logits for multi-label classification.

In our model, the embedding dimension is $H = 64$ and the number of classes is $C = 5$.

Convolutional blocks extract local waveform features and reduce the temporal dimension by scanning the signal with small filters, stabilizing activations via batch normalization, injecting non-linearity with ReLU, and halving the time length through max-pooling. *GRU layer* integrates these features over time, retaining important patterns and discarding noise. *RealNVP flow* warps the embedding distribution into a flexible latent space by alternately scaling and shifting parts of the vector under learned transformations. *MLP head* maps the latent vector to class logits, which are converted to independent probabilities by sigmoid at inference.

4.4 Training

We trained the network with Adam (LR= 10^{-4}) for up to 1000 epochs, applying early stopping after 25 non-improving epochs and gradient norm clipping to prevent exploding updates. This setup ensured fast convergence while avoiding overfitting, as evidenced by the loss curves in Figure A1. An error-pattern analysis (Table A1) highlights the most frequent confusions and guides future refinements. For full experimental details and additional plots, see Appendix A.1.

4.5 Threshold selection

To balance precision and recall across imbalanced classes, we determined optimal decision thresholds on the validation set by sweeping probabilities to maximize each class’s F1 score. Adopting these tailored cut-offs improved the model’s overlap measures (notably Jaccard and F1) on test data while only slightly reducing exact-match accuracy (see Figure A2). Detailed methodology and threshold values are available in Appendix A.2.

4.6 SHAP value computation

We used the SHAP library² (version 0.47.0) and its *GradientExplainer* to attribute model predictions to input features. *GradientExplainer* combines model gradients with a background distribution to efficiently approximate Shapley values. A representative subset (fold 8) of the data served as the background distribution. For each batch of ECG segments (shape $(B, 1000, 12)$), the explainer produced raw SHAP values of shape $(B, 1000, 12, 5)$ - one value per time step, channel, and output class - which were then transposed to $(B, 5, 1000, 12)$ for consistency with our downstream analysis. To handle large test sets efficiently, SHAP values were written incrementally into an HDF5 dataset, allowing interrupted runs to resume without loss.

4.6.1 Rule extraction from SHAP explanations

The following steps follow the *PHAR* (Mozolewski et al., 2025) methodology for converting numeric attributions into interval rules.

To ensure that generated rules reflect the model’s actual decisions, we first convert the model’s logits into binary predictions using the previously determined per-class thresholds. We then iterate over each class that the model predicts as present, slicing out its corresponding SHAP explanation matrix of shape (N, T, M) from the full (N, C, T, M) array. Rules are only extracted for these positively predicted classes.

Important features - the ones which will be included in the rules for given data instance - are chosen by applying a single global cutoff at the specified percentile of all absolute SHAP values. In our implementation we used *percentile* = 90.0.

For each selected time-channel coordinate, we generate 1000 perturbed samples by adding Gaussian noise scaled by $\sigma = 1.0$. We enforce that the entire multi-label output remains unchanged under these perturbations. Enforcing it means that when we derive a rule for a particular instance and class, we only accept intervals for which every perturbed version of that instance yields the exact same full set of predicted labels. In other words, the rule for one class cannot flip any other class’s prediction on that example - so if an instance is assigned to multiple classes, each class’s rule will remain compatible and never contradict the others.

The maximal range of values that preserve consistency defines a simple interval rule (e.g. “time_t_chan_m $\in$ (a,b]”). Finally, we evaluate each rule’s coverage (fraction of samples satisfying the interval) and confidence (fraction of those whose full prediction vector matches the original). The result is a compact set of interval rules, each annotated with its empirical support and reliability.

4.7 Prototype Selection Implementation

We implemented the prototype selection pipeline on the 2,198 samples from an unused fold of the data. The process consisted of the following steps:

²<https://shap.readthedocs.io/en/latest/>

4.7.1 Sample Filtering

From the initial dataset, we retained only samples that met two criteria: (i) correct classification by our trained model, and (ii) single-label assignment. This filtering resulted in 947 samples distributed as: NORM (377/380, 99.21%), MI (160/317, 50.47%), STTC (231/537, 43.02%), CD (156/330, 47.27%), and HYP (23/845, 2.72%). Overall retention was 947/2,198 (43.08%). (Note: the sum of the per-class denominators is 2,409, but only 2,198 unique samples existed due to multi-class overlaps). The filtering ensures prototype quality by excluding ambiguous or potentially mislabeled cases.

4.7.2 DTW Distance Computation

For each diagnostic class, we computed pairwise DTW distances between all samples. We used the *similarity-ts*³ package implementation with the *DTW* metric and a stride of 1 to reduce computational complexity while maintaining accuracy. The DTW distance matrices captured morphological similarities between ECG waveforms within each class. Distance computation took approximately 89 hours for all classes on a system with 64 CPU cores, however, the code was not specifically parallelized, and the libraries used (per their documentation) do not utilize multithreading.

4.7.3 MDS Projection and Clustering

We applied Multi-Dimensional Scaling to each class’s DTW distance matrix. The optimal number of MDS components was determined using silhouette analysis:

- NORM: 6 components (silhouette score: 0.44)
- MI: 7 components (silhouette score: 0.32)
- STTC: 6 components (silhouette score: 0.36)
- CD: 5 components (silhouette score: 0.032)
- HYP: 3 components (silhouette score: 0.28)

For k-means clustering in MDS space, we evaluated $k \in \{2, 3, \dots, 10\}$ using combined metrics of silhouette coefficient and the optimal cluster counts were: NORM ($k=6$), MI ($k=7$), STTC ($k=6$), CD ($k=5$), and HYP ($k=3$). Figure 3 illustrates this process for the CD class, where silhouette analysis revealed $k=5$ as optimal. The resulting clustering, visualized in Figure 4, shows separation of CD samples into four distinct morphological groups

4.7.4 Medoid Extraction

From each cluster, we identified the medoid as the sample with the minimum average DTW distance to all other cluster members. As shown in Figure 4 for the CD class, medoids (marked with red circles) represent the most central examples within each cluster, ensuring representativeness. This resulted in 27 total prototypes across all classes, stored as a prototype database for counterfactual generation close to real-time.

³ Available at <https://github.com/alejandrofdez-us/similarity-ts>

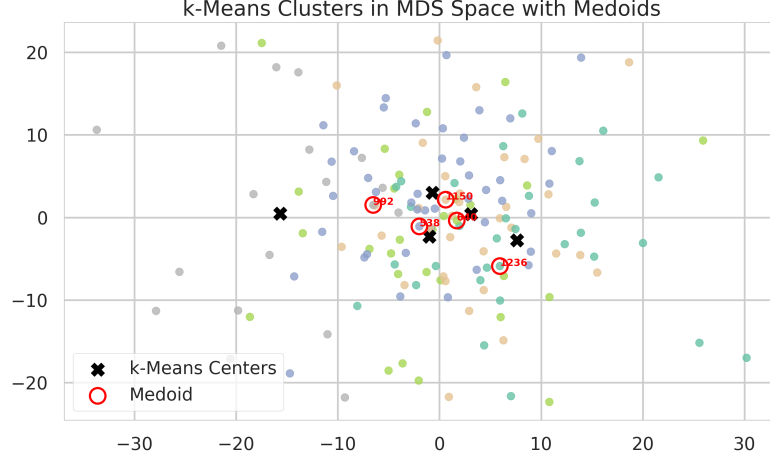


Fig. 4 k-Means clustering results in MDS space for Conduction Disturbance (CD) class.

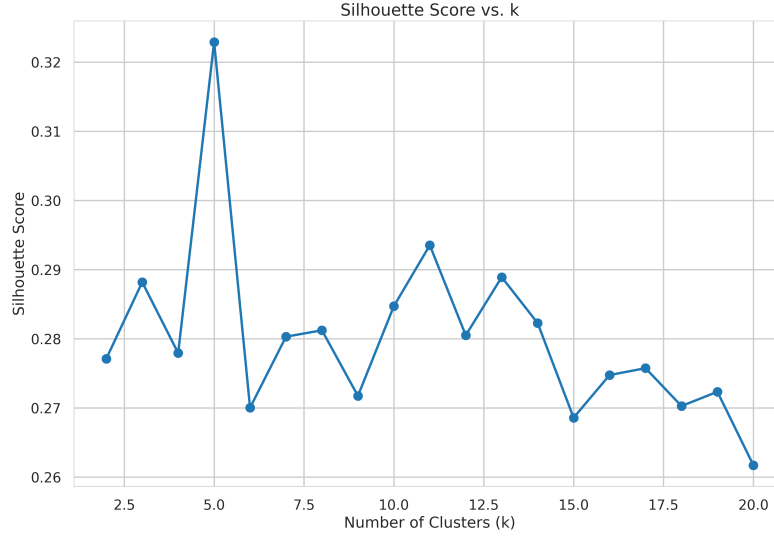


Fig. 5 Silhouette analysis for optimal cluster selection in the Conduction Disturbance (CD) class.

5 Results and Discussion

5.1 Experimental Results

This section presents a comprehensive evaluation of our prototype-driven counterfactual generation framework across three distinct approaches: *Original* (prototype without optimization), *Sparse* (prototype with sparsity optimization), and *Aligned Sparse* (prototype with both R-peak alignment and sparsity optimization). Our evaluation employs five key metrics specifically adapted for multivariate time-series

ECG data: *validity*, *sparsity* (L_0, L_1, L_2), *stability* (noise and temporal), and *decision margin*.

5.1.1 Overall Performance Analysis

Table 3 Original vs counterfactuals: summary statistics of validity, sparsity, stability, and decision margin.

Metric	Original	Sparse	Aligned Sparse
Validity_{multi}	1.0000	1.0000	0.8126
Sparsity			
Sparsity ratio	0.9837	0.7957	0.7800
L_0 sparsity	0.9837	0.7957	0.7800
L_1 sparsity	0.2868	0.2700	0.2018
L_2 sparsity	0.2512	0.2442	0.1749
Stability			
Noise stability	1.0000	1.0000	0.9948
Temporal stability	0.5304	0.5097	0.7593
Margin	0.7800	0.7750	0.6933

Bold values indicate best performance per metric. For sparsity metrics, lower values are better (sparser counterfactuals). For all other metrics, higher values are better.

Table 3 presents the aggregate performance across all 963 test samples, revealing distinct trade-offs between the three approaches. The *validity* metric, fundamental to any counterfactual explanation, demonstrates perfect performance (1.0) for both *Original* and *Sparse* methods, which means generated counterfactuals successfully change the model’s prediction to the desired target class. However, the *Aligned Sparse* method shows a notable validity drop to 0.8126, representing an 18.74% decrease. This reduction stems from the temporal alignment process, which occasionally produces waveforms that, while physiologically more plausible, fail to cross the model’s decision boundaries for certain classes. For example, when a prototype has fewer beats than the original signal, we prepend zeros to its start to allow a clear overlay on the query sample, but this may negatively affect the counterfactual’s validity.

Sparsity performance reveals the effectiveness of our optimization approach. The sparsity ratio improves significantly from 0.9837 (Original) to 0.7800 (Aligned Sparse), indicating that our method modifies only 78% of the original signal rather than nearly the entire waveform. This 20.7% improvement in sparsity directly translates to enhanced interpretability, as clinicians can focus on specific ECG segments rather than examining wholesale signal replacements. The L_1 and L_2 sparsity metrics show consistent improvements, with L_2 sparsity decreasing from 0.2512 to 0.1749, demonstrating that modifications are not only fewer but also smaller in magnitude.

Stability analysis reveals complementary strengths across methods. Noise stability remains consistently high (≥ 0.9948) across all approaches, indicating robust performance under measurement noise typical in clinical ECG recordings. Also, temporal

stability shows a dramatic improvement from 0.5304 (Original) to 0.7593 (Aligned Sparse), representing a 43.2% enhancement. This improvement is clinically significant, as it indicates that counterfactuals maintain their validity even when subjected to small temporal shifts common in ECG acquisition.

The *decision margin* metric shows a trade-off between sparsity and confidence. While the *Original* method achieves a margin of 0.7800, the Aligned Sparse approach shows a decrease to 0.6933. Despite this 11.1% reduction, the margin remains well above clinical acceptability thresholds, suggesting that the trade-off favoring interpretability and temporal coherence is justified.

5.1.2 Class-Specific Performance Analysis

Table 4 Comparison of counterfactual generation methods by metric and target class.

Metric	Method	CD	HYP	MI	NORM	STTC
Validity _{multi}	Original	1.000	1.000	1.000	1.000	1.000
	Sparse	1.000	1.000	1.000	1.000	1.000
	Aligned Sparse	0.798	0.132	0.989	0.681	0.650
L_0 sparsity	Original	0.985	0.982	0.982	0.987	0.984
	Sparse	0.810	0.797	0.747	0.940	0.748
	Aligned Sparse	0.775	0.828	0.703	0.914	0.817
Noise stability	Original	1.000	1.000	1.000	1.000	1.000
	Sparse	1.000	1.000	1.000	1.000	1.000
	Aligned Sparse	1.000	0.978	1.000	0.991	0.987
Margin	Original	0.607	0.446	0.954	0.695	0.598
	Sparse	0.605	0.445	0.946	0.691	0.597
	Aligned Sparse	0.472	0.311	0.900	0.467	0.616

Bold values indicate best performance for each target class-metric combination. For L_0 sparsity, lower values are better (sparser counterfactuals). For all other metrics, higher values are better.

Table 4 reveals significant performance variations across diagnostic classes, providing insights into the inherent challenges of different conditions. Myocardial Infarction (MI) demonstrates the strongest overall performance, with *validity* of 0.989 and superior sparsity (L_0 : 0.703) in the *Aligned Sparse* condition. This strong performance aligns with clinical expectations, as MI typically presents with distinct ECG features.

Conduction Disturbances (CD) show robust **validity** (0.798) but require more extensive modifications (L_0 : 0.775), reflecting the global nature of conduction abnormalities that affect timing relationships across multiple ECG leads.

Hypertrophy (HYP) presents the most significant challenge, with *validity* dropping to only 0.132 in the Aligned Sparse condition. This poor performance reflects several factors: (i) HYP represents the smallest class in our dataset (12.15%), this imbalance leads to limited prototype diversity; (ii) the condition often co-occurs with other pathologies, creating complex multi-label scenarios.

Decision margin analysis by class reveals MI counterfactuals maintain the highest confidence (0.9), while HYP shows the lowest (0.311). This pattern correlates strongly with *validity* performance, pointing out that classes with poor validity also suffer from low confidence in successful cases.

5.2 Advantages of the prototype-based counterfactuals

In a clinical context, our prototype-based real-time explanation framework for ECG delivers three concrete benefits:

- **Computational efficiency:** By clustering rule sets into k representative prototypes, inference requires only k prototype comparisons instead of N individual-instance evaluations, reducing complexity from $\mathcal{O}(N)$ to $\mathcal{O}(k)$, where $k \ll N$.
- **Clinical validity:** Using actual ECG samples as prototypes guarantees physiologically plausible waveforms, avoiding the synthetic artifacts that can arise from interpolation or generation methods.
- **Interpretability:** Each prototype represents a cluster of similar cases, allowing clinicians to understand not just individual predictions but also the typical patterns within each diagnostic category.

Medoid-based selection enables our prototypes to capture within-class morphological diversity by choosing the most centrally located ECG in each cluster, ensuring both coverage and representativeness. This balance is crucial for producing concise, clinically meaningful counterfactual exemplars across heterogeneous patient populations.

5.3 Explanation Representation and Expert Feedback

While technical performance metrics provide essential validation of our approach, understanding healthcare professionals’ preferences and cognitive needs is equally critical for successful clinical deployment. The representation and presentation of explanations can significantly impact their utility, regardless of underlying technical quality. To address this challenge, we conducted semi-structured interviews with three practitioners spanning 3 to 40+ years of ECG interpretation experience and varying levels of familiarity with AI systems. Each expert was presented with different visualization modes, including lead-wise importance plots, attention span highlights, and single vs. multiple counterfactual overlays, for identical ECG inputs. They evaluated these modes based on interpretability, cognitive load, clinical relevance, and workflow compatibility.

All experts strongly favored single counterfactual visualizations paired with attention highlighting, citing reduced cognitive effort and improved clarity in identifying diagnostic shifts. Rather than viewing all possible class transitions, they preferred focused, user-driven alternatives relevant to specific clinical scenarios. Experts emphasized the importance of aligning lead-level explanations with domain knowledge and voiced concerns when model attributions diverged from clinically expected leads.

One of the experts provided a valuable insight: the granularity of explanation should be adapted to its clinical or research context. In investigative settings, detailed

probabilistic confidence scores and channel-weight attributions help uncover novel ECG patterns, while in everyday clinical practice, a simple binary indication of importance proves both sufficient and more user-friendly. The expert further advised tailoring visualization scope by care environment: on general wards, highlighting a single representative beat for non-urgent cases; in emergency departments, offering a concise rationale for deviation from normal to accelerate triage decisions; and in specialized outpatient clinics, presenting richer counterfactual scenarios to aid in distinguishing disease subtypes and supporting complex diagnostic reasoning.

These findings underscore the need for interactive, clinician-controlled explanation systems. Experts expressed a clear preference for adaptive visualizations over static, exhaustive outputs and highlighted integration with diagnostic workflows as essential. Their feedback directly informed our decision to prioritize an interactive platform allowing case-specific exploration, real-time adjustment of explanation complexity, and seamless integration with clinical tools.

5.4 Limitations and Challenges

Several limitations emerge from our analysis. The *validity-sparsity trade-off* in the aligned method represents a fundamental challenge of achieving physiologically realistic counterfactuals while maintaining model prediction changes. This trade-off is particularly pronounced for HYP, suggesting that certain cardiac conditions may be inherently difficult to explain through prototype-based approaches. Class imbalance effects significantly impact performance, with HYP (12.15% of the dataset) showing markedly worse performance than well-represented classes like NORM (43.64%). Future work should investigate targeted sampling strategies and class-aware prototype selection to address these disparities.

The *computational complexity* of the alignment process, while acceptable for our current implementation, may limit real-time deployment in resource-constrained clinical environments. The DTW-based alignment requires $O(n^2)$ operations per prototype comparison, potentially limiting scalability to larger prototype databases.

Evaluation scope remains limited to the PTB-XL dataset, and generalization to other ECG datasets, different demographics, or varying acquisition protocols requires further validation. Additionally, our expert evaluation, while promising, requires expansion to larger clinical studies to establish definitive usability and accuracy benchmarks.

5.5 Future Improvements

The results presented in Section 5.1, demonstrate significant potential for deployment in clinical ECG interpretation systems. The high validity rates for common conditions (MI, CD) and the improved interpretability through sparse modifications address key requirements for AI-assisted diagnosis. The framework’s ability to generate explanations in near real-time makes it suitable for point-of-care applications.

Future improvements should focus on *adaptive validity thresholds* that balance explanation quality with clinical requirements for different conditions. For highly critical conditions, perfect validity may be essential. *Hybrid prototype selection strategies*

could address class imbalance by combining data-driven clustering with expert-curated examples for rare conditions. This approach would ensure that prototype databases contain representative examples even for uncommon pathologies. And also, *real-time adaptation* mechanisms could allow the system to learn from clinician feedback, gradually improving prototype selection and counterfactual generation based on expert preferences and clinical outcomes.

The metrics used in our evaluation process capture distinct aspects of counterfactual quality. Similar to [Darias et al. \(2025\)](#) and [S. Wang, Bates, Aronow, and Jordan \(2023\)](#), the natural next step is to combine them into a single composite quality metric Q that aggregates four key criteria: validity, sparsity (via the L_0 ratio), stability, and decision margin. ($Q(\mathbf{x}') = w_v \cdot \text{validity} + w_s \cdot (1 - \text{sparsity_ratio}) + w_{st} \cdot \text{stability_score} + w_m \cdot \text{decision_margin}$) The main challenge lies in determining the optimal weights (w_v, w_s, w_{st}, w_m), where each weight respectively governs the contribution of each of the aforementioned metrics, for a given clinical application. In future work, we will employ systematic grid search and Bayesian optimization to select weights that best align Q with clinician-assessed explanation quality. Furthermore, integrating a human(experts)-in-the-loop feedback pipeline will allow iterative, live fine-tuning of these weights based on real-world clinical validation.

Our current overlays on ECG plots use SHAP values. To enhance the contrast and readability of ECG feature highlights, we will investigate alternative numeric attribution methods beyond SHAP ([Lundberg & Lee, 2017](#)), such as LIME ([Ribeiro et al., 2016](#)), Integrated Gradients ([Sundararajan, Taly, & Yan, 2017](#)), DeepLIFT ([Shrikumar, Greenside, & Kundaje, 2017](#)), and SmoothGrad ([Smilkov, Thorat, Kim, Viégas, & Wattenberg, 2017](#)). We also plan to explore non-linear transformations of SHAP values or thresholds to enhance contrast and clarity in the visual highlights.

The development of *interactive explanation platform* will allow clinicians to explore different counterfactual scenarios, adjust sparsity levels, and visualize explanation sensitivity to various modifications. Such platforms could transform counterfactual explanations from static outputs to dynamic diagnostic tools.

6 Conclusions

Electrocardiogram (ECG) interpretation is essential for diagnosing cardiac disorders under urgent clinical conditions. Modern deep learning classifiers achieve remarkable accuracy by extracting complex features from large ECG datasets, but they operate as opaque models (a.k.a. black boxes). This lack of transparency undermines clinician confidence and obstructs their adoption in everyday practice, where understanding the reasoning process behind a decision is mandatory. Consequently, delivering trustworthy, physiology-based explanations in real time is critical to support rapid, evidence-driven patient care. Especially useful in this situation are counterfactual explanations, because they reveal minimal and plausible changes needed to alter ECG classification.

This work makes five key contributions. First, we validate counterfactual and heatmap visualizations with domain experts in ECG interpretation. Second, we propose a prototype-based selection method to select representative counterfactual

examples of the entire data distribution. Third, we develop alignment and sparsification optimization techniques to map and prune prototypes, yielding clear, query-specific explanations. Fourth, our approach generalizes beyond ECG to other multivariate time-series domains, such as wearable patient vitals, industrial sensor networks, or predictive maintenance systems, since its core components remain model-agnostic and can be adapted with modest domain-specific alignment. Fifth, we adopt evaluation metrics for counterfactual explanations, originally developed in the XAI literature, and tailor them to the unique characteristics of multivariate time-series ECG signals, ensuring meaningful assessment of our method’s fidelity.

Expert interviews with ECG specialists confirmed that our counterfactual visualizations and heatmaps improve clinical reasoning and align with physiological knowledge. We applied a novel *Post-hoc Attribution Rule extraction (PHAR)* method to extract logical rules for each ECG sample, capturing key morphological features such as amplitude thresholds. From these rules, we defined a distance metric that measures similarity between ECG waveforms while preserving physiological plausibility. Clustering with medoids yielded representative prototypes from non-query classes, which served as counterfactual examples. A dedicated alignment algorithm then mapped each prototype onto the query sample’s temporal structure. Prototype sparsification produced more concise, targeted explanations. Quantitative evaluation on validity, sparsity, stability, and margin metrics confirmed excellent validity for both *Original* and *Sparse* counterfactuals, with only minor degradation for sparse aligned variants; margins remained clear and stability high across all types.

To conclude, our explainable ECG classification framework unites high accuracy with real-time transparency. Clinicians can potentially receive intuitive, counterfactual insights that foster trust and support rapid decision-making. Moreover, its largely model-agnostic design extends naturally to diverse multivariate time-series data—enabling applications from wearable patient vitals to industrial sensor monitoring and predictive maintenance.

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

Declarations

Ethics approval and consent to participate. This study involved two components. First, it used public, de-identified ECG datasets. Second, it included semi-structured interviews with domain experts (ECG specialists). All expert participants were adults, participated voluntarily, and provided informed consent before taking part. The activity collected only professional opinions about methods and clinical workflows. No personal data or patient information were collected. Responses were recorded and analysed in anonymized, aggregated form. In accordance with the authors’ institutional guidelines, this expert consultation did not require formal approval by a Research Ethics Committee because it did not involve patients, vulnerable individuals, or identifiable personal data.

Consent for publication. All expert participants consented to the publication of anonymized and aggregated results. No identifiable quotations or personal information are presented.

Availability of data and material. This study uses only public, de-identified ECG datasets. The PTB-XL dataset is publicly available from community repositories listed in the References. No new patient data were collected by the authors.

The code used for data processing, model training, explanation generation, and figure creation is available from the authors upon reasonable request for academic use. Trained model weights and auxiliary artifacts (e.g., SHAP attributions and prototype identifiers) can also be shared upon request. To protect participant anonymity, raw interview notes and individual survey responses are not publicly shared. The interview protocol is available from the authors upon reasonable request for academic research.

There are no proprietary restrictions that prevent sharing of the materials. There are no license or contractual restrictions that prevent sharing for academic research.

Competing interests. The authors declare no competing interests.

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 No. 101120406. The paper reflects only the authors’ view and the European Commission is not responsible for any use that may be made of the information contained herein.

This work was supported by the ELLIIT Excellence Center at Linköping–Lund in Information Technology in Sweden.

Contribution of Maciej Mozolewski 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.

This publication has been funded by the SFI NorwAI (Centre for Research-based Innovation, 309834). Betül Bayrak gratefully acknowledges the financial support from the Research Council of Norway and the partners of the SFI NorwAI.

Author contributions.

Maciej Mozolewski: Prepared and preprocessed the ECG dataset, including data splitting, signal normalization, and background sample preparation for SHAP analysis. Implemented the SHAP attribution pipeline using GradientExplainer with efficient HDF5-based storage. Designed and developed the deep ECG classifier with an advanced training regimen and threshold calibration based on F1 score. Adapted the *PHAR* rule-extraction module for the multiclass setting, enabling overlapping class assignments per instance. Integrated R-peak detection and prototype temporal alignment. Executed computational analyses, collated the resulting metrics and figures, and selected representative ECG signals for visualization. Automated the end-to-end data workflow for comprehensive table and figure generation. Drafted, structured, and unified the Introduction and Background sections of the manuscript.

Betül Bayrak: Conceptualization of the counterfactual explanation generation approach. Development of prototype-guided counterfactuals, including implementation of DTW distance computation and clustering, prototype extraction, counterfactual selection, and sparsity optimization method. Development of quantitative evaluation metrics, which are validity, sparsity, stability, and margin metrics, and implementation of the query processing workflow. Organization and execution of cardiologist interviews, survey design, and data collection; creation of workflow and architecture figures.

Maciej Mozolewski and Betül Bayrak: Conceptualization of the study and manuscript framework; interpretation of results; creation of data visualizations; drafting of Materials and Methods, Results and Discussion, and Conclusions; final review and approval of the manuscript.

Grzegorz J. Nalepa: Provided critical feedback on content and structure. Ensured clarity, consistency, and accuracy of language. Edited and refined the final manuscript.

Kerstin Bach: Facilitated interviews with ECG experts by identifying suitable experts and making introductions. Advised on interview scope and survey materials. Reviewed the final manuscript.

References

- Ao, R., & He, G. (2023). Image based deep learning in 12-lead ecg diagnosis. *Frontiers in Artificial Intelligence, Volume 5 - 2022*, , <https://doi.org/10.3389/frai.2022.1087370> Retrieved from <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.1087370>
- Aryal, S., & Keane, M.T. (2023, 8). Even if explanations: Prior work, desiderata & benchmarks for semi-factual xai. E. Elkind (Ed.), *Proceedings of the thirty-second international joint conference on artificial intelligence, IJCAI-23* (pp. 6526–6535). International Joint Conferences on Artificial Intelligence Organization. Retrieved from <https://doi.org/10.24963/ijcai.2023/732> (Survey Track)
- Baig, Z., Nasir, S., Khan, R.A., Haque, M.Z.U. (2025). *Arrhythmia-vision: Resource-conscious deep learning models with visual explanations for ecg arrhythmia classification*. Retrieved from <https://arxiv.org/abs/2505.03787>
- Bayrak, B., & Bach, K. (2023). Pertcf: A perturbation-based counterfactual generation approach. *International conference on innovative techniques and applications of artificial intelligence* (pp. 174–187).
- Bayrak, B., & Bach, K. (2024). Evaluation of instance-based explanations: an in-depth analysis of counterfactual evaluation metrics, challenges, and the ceval toolkit. *IEEE Access*, ,

- Darias, J.M., Bayrak, B., Caro-Martínez, M., Díaz-Agudo, B., Recio-Garcia, J.A. (2024). An empirical analysis of user preferences regarding xai metrics. *International conference on case-based reasoning* (pp. 96–110).
- Darias, J.M., Díaz-Agudo, B., Recio-Garcia, J.A. (2025). Evaluating objective metrics for time series model explainability. *International conference on case-based reasoning* (pp. 189–203).
- Delaney, E., Greene, D., Keane, M.T. (2021). Instance-based counterfactual explanations for time series classification. A.A. Sánchez-Ruiz & M.W. Floyd (Eds.), *Case-based reasoning research and development* (pp. 32–47). Cham: Springer International Publishing.
- Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., ... Stanley, H. (2000, 07). Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals. *Circulation*, 101, E215-20, <https://doi.org/10.1161/01.CIR.101.23.e215>
- Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., Ng, A.Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65–69, <https://doi.org/10.1038/s41591-018-0268-3> Retrieved from <https://doi.org/10.1038/s41591-018-0268-3> (Accessed on 2019-01-01)
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312, <https://doi.org/10.1002/widm.1312>
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), e1312, <https://doi.org/https://doi.org/10.1002/widm.1312> Retrieved from <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1312> <https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312>
- Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems* (p. 4768–4777). Red Hook, NY, USA: Curran Associates Inc.
- Mothilal, R.K., Sharma, A., Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607–617).

- Mozolewski, M., Bobek, S., Nalepa, G.J. (2025). Explaining time series classifiers with phar: Rule extraction and fusion from post-hoc attributions. *arXiv preprint*, , Retrieved from <https://arxiv.org/abs/2508.01687> (Submitted) [arXiv:2508.01687](https://arxiv.org/abs/2508.01687) [cs.LG]
- Nayebi, A., Tipirneni, S., Reddy, C.K., Foreman, B., Subbian, V. (2023). Windowshap: An efficient framework for explaining time-series classifiers based on shapley values. *Journal of Biomedical Informatics*, 144, 104438, <https://doi.org/https://doi.org/10.1016/j.jbi.2023.104438> Retrieved from <https://www.sciencedirect.com/science/article/pii/S1532046423001594>
- Olszewski, R.T., Maxion, R., Siewiorek, D. (2001). *Generalized feature extraction for structural pattern recognition in time-series data* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA, USA. (AAI3040489)
- Proniewska, K.K., Abächerli, R., van Dam, P.M. (2023). The Δ waveecg: The differences to the normal 12-lead ecg amplitudes. *Journal of Electrocardiology*, 76, 45-54, <https://doi.org/https://doi.org/10.1016/j.jelectrocard.2022.10.014> Retrieved from <https://www.sciencedirect.com/science/article/pii/S0022073622002138>
- Ribeiro, M.T., Singh, S., 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* (p. 1135–1144). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/2939672.2939778>
- Röber, T.E., Goedhart, R., Birbil, Ş.İ. (2025). *Clinicians' voice: Fundamental considerations for xai in healthcare*. Retrieved from <https://arxiv.org/abs/2411.04855>
- Röber, T.E., Lumadjeng, A.C., Akyüz, M.H., Birbil, Ş.İ. (2025). Rule generation for classification: Scalability, interpretability, and fairness. *Computers & Operations Research*, 183, 107163, <https://doi.org/https://doi.org/10.1016/j.cor.2025.107163> Retrieved from <https://www.sciencedirect.com/science/article/pii/S0305054825001911>
- Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 43-49, <https://doi.org/10.1109/TASSP.1978.1163055>
- Salih, A.M., Galazzo, I.B., Gkontra, P., Rauseo, E., Lee, A.M., Lekadir, K., ... Menegaz, G. (2024, 08). A review of evaluation approaches for explainable ai with applications in cardiology. *Artificial Intelligence Review*,

57(9), 240, <https://doi.org/10.1007/s10462-024-10852-w> Retrieved from <https://doi.org/10.1007/s10462-024-10852-w> (Epub 2024 Aug 9)

- Shrikumar, A., Greenside, P., Kundaje, A. (2017). Learning important features through propagating activation differences. *Proceedings of the 34th international conference on machine learning - volume 70* (p. 3145–3153). JMLR.org.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, ,
- Strodthoff, N., Mehari, T., Nagel, C., Aston, P.J., Sundar, A., Graff, C., . . . Schaeffter, T. (2023). Ptb-xl+, a comprehensive electrocardiographic feature dataset. *Scientific Data*, 10(1), 279, <https://doi.org/10.1038/s41597-023-02153-8> Retrieved from <https://doi.org/10.1038/s41597-023-02153-8>
- Sundararajan, M., Taly, A., Yan, Q. (2017, 06–11 Aug). Axiomatic attribution for deep networks. D. Precup & Y.W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 3319–3328). PMLR. Retrieved from <https://proceedings.mlr.press/v70/sundararajan17a.html>
- Tanyel, T., Atmaca, S., Gökçe, K., Balık, M.Y., Güler, A., Aslanger, E., İlkay Öksüz. (2025). Interpretable ecg analysis for myocardial infarction detection through counterfactuals. *Biomedical Signal Processing and Control*, 102, 107227, <https://doi.org/https://doi.org/10.1016/j.bspc.2024.107227> Retrieved from <https://www.sciencedirect.com/science/article/pii/S1746809424012850>
- Theissler, A., Spinnato, F., Schlegel, U., Guidotti, R. (2022). Explainable ai for time series classification: A review, taxonomy and research directions. *IEEE Access*, 10, 100700–100724, <https://doi.org/10.1109/ACCESS.2022.3207765>
- van Dam, P.M., Boonstra, M., Locati, E.T., Loh, P. (2021). The relation of 12 lead ecg to the cardiac anatomy: The normal cineecg. *Journal of Electrocardiology*, 69, 67–74, <https://doi.org/https://doi.org/10.1016/j.jelectrocard.2021.07.014> Retrieved from <https://www.sciencedirect.com/science/article/pii/S0022073621001576>
- Van Looveren, A., & Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, & J.A. Lozano (Eds.), *Machine learning and knowledge discovery in databases. research track* (pp. 650–665). Cham: Springer International Publishing.

- Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., Schaeffter, T. (2022). *PTB-XL, a large publicly available electrocardiography dataset*. PhysioNet dataset, version 1.0.3, <https://physionet.org/content/ptb-xl/1.0.3/>. PhysioNet. Retrieved from <https://physionet.org/content/ptb-xl/1.0.3/> (RRID:SCR_007345)
- Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T. (2020). Ptb-xl: A large publicly available ecg dataset. *Scientific Data*, 7, 191, <https://doi.org/10.1038/s41597-020-0495-6>
- Wang, F., Jiang, Y., Zhang, R., Wei, A., Xie, J., Pang, X. (2025). A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions. *Sensors*, 25(1), , <https://doi.org/10.3390/s25010190> Retrieved from <https://www.mdpi.com/1424-8220/25/1/190>
- Wang, S., Bates, S., Aronow, P., Jordan, M.I. (2023). Operationalizing counterfactual metrics: Incentives, ranking, and information asymmetry. *arXiv preprint arXiv:2305.14595*, ,
- Wondimu, N.A., Buche, C., Visser, U. (2022). Interactive machine learning: A state of the art review. *ArXiv*, *abs/2207.06196*, , Retrieved from <https://api.semanticscholar.org/CorpusID:250491626>

Appendix A Details of model training and threshold selection

A.1 Training

Training optimizes the multi-label binary cross-entropy loss using the *Adam optimizer* with a constant learning rate of 10^{-4} . The model runs for up to 1000 epochs, with training and validation losses recorded each epoch. Early stopping halts training if the validation loss does not improve for 25 consecutive epochs, restoring the best model state. To maintain stability, gradient norm clipping is applied throughout training. Figure A1 displays the training and validation losses over epochs. Training loss steadily

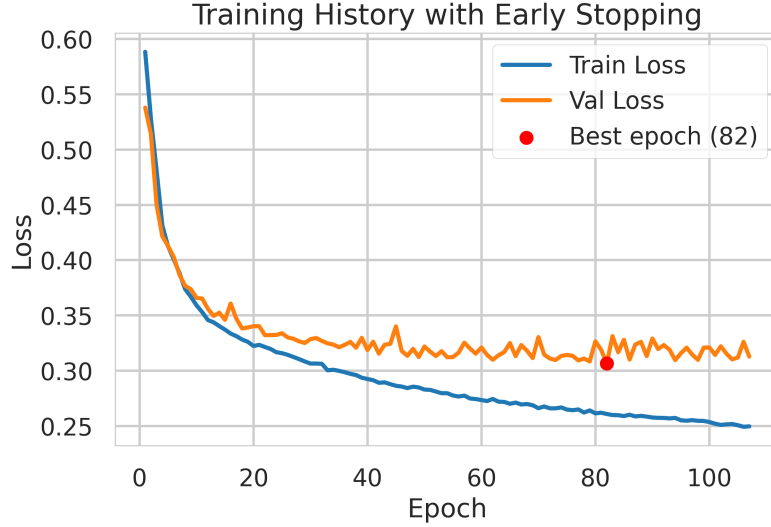


Fig. A1 Training and validation loss with early stopping.

decreases, while validation loss plateaus around epoch 80, at which point early stopping was triggered.

Evaluation on the test set using a uniform threshold of 0.5 for all classes yielded a Hamming loss of 0.1285, corresponding to an exact-match accuracy of 87.15%. The Jaccard score reached 0.6754, while the F1 metrics were 0.7359 (micro), 0.6973 (macro) and 0.7295 (weighted). No class-specific thresholds were applied in this evaluation.

To understand where the model still struggles, we conducted an error-pattern analysis and collected all misclassifications affecting at least 1% of the test samples (see Table A1). This breakdown reveals, for example, that myocardial infarction cases are most often confused with normal ECGs, and that mixed labels (e.g. CD + NORM) sometimes collapse to single-class predictions.

Table A1 Misclassification patterns above 1% of test samples

True	Predicted	% of test
MI	NORM	3.14
NORM	—	2.46
CD	NORM	1.77
STTC	NORM	1.55
CD, NORM	NORM	1.32
NORM	STTC	1.32
STTC	—	1.27
HYP	NORM	1.23
CD	CD, MI	1.14
CD, MI	CD	1.05
NORM	MI	1.00

Abbreviations: NORM – Normal ECG; MI – Myocardial Infarction; STTC – ST/T Change; CD – Conduction Disturbance; HYP – Hypertrophy.

A.2 Threshold selection

Optimal decision thresholds for each class were found by scanning the validation set probabilities from 0.0 to 1.0 in steps of 0.001 and choosing the value that maximized the binary F1 score for that class. This procedure yielded thresholds of approximately 0.307 for normal ECG (*NORM*), 0.316 for myocardial infarction (*MI*), 0.336 for conduction disturbance (*CD*), 0.352 for ST/T change (*STTC*), and 0.446 for hypertrophy (*HYP*).

Applying these class-specific thresholds on the test set produced a Hamming loss of 0.1381 (accuracy 86.19%), Jaccard score 0.6935, F1 micro 0.7436, F1 macro 0.7094 and F1 weighted 0.7419, showing a slight trade-off in exact matches but improved overlap and balanced F1 across classes.

Figure A2 plots the true positive rate against the false positive rate for each class as the classification threshold varies. The area under each curve (AUC) quantifies how well the model separates positive from negative examples: ST/T Change (STTC) achieves the highest AUC (0.94), indicating very good discrimination, followed by Hypertrophy (HYP) at 0.92. The normal (NORM), infarction (MI) and conduction disturbance (CD) classes all reach $AUC \approx 0.89$, demonstrating strong but slightly lower separability.

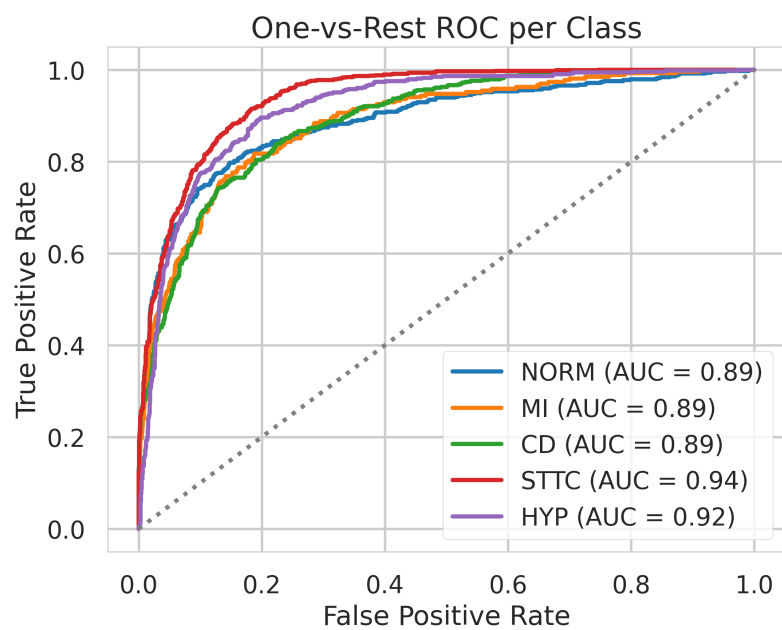


Fig. A2 One-versus-rest ROC curves per class on validation data.

Appendix B Metrics for ECG counterfactuals

In this appendix, we report mean values of four counterfactual quality metrics, *multi-class validity* ($Validity_{multi}$), L_0 sparsity, *noise stability*, and *decision margin* for each initial ECG class. For each class, three series are evaluated: the *Original* ECG signal, the *Sparse* counterfactual before alignment, and the *Aligned Sparse* counterfactual after temporal alignment. Tables B2–B6 list these metrics for the five single classes (NORM, STTC, CD, MI, HYP), while Table B7 and Table B8 presents the same metrics for two-class combinations.

Table B2 Mean counterfactual metrics for initial class NORM (n=564)

Target Class	Series	$Validity_{multi}$	L_0 sparsity	Noise stability	Margin
CD	Original	1.0000	0.9871	1.0000	0.6080
	Sparse	1.0000	0.8114	1.0000	0.6055
	Aligned Sparse	0.8144	0.7584	1.0000	0.4632
HYP	Original	1.0000	0.9822	1.0000	0.4499
	Sparse	1.0000	0.7892	1.0000	0.4495
	Aligned Sparse	0.1163	0.8396	1.0000	0.2815
MI	Original	1.0000	0.9835	1.0000	0.9591
	Sparse	1.0000	0.7545	1.0000	0.9508
	Aligned Sparse	0.9929	0.6942	1.0000	0.9062
STTC	Original	1.0000	0.9856	1.0000	0.5932
	Sparse	1.0000	0.7442	1.0000	0.5927
	Aligned Sparse	0.6162	0.8124	0.9891	0.6176

Table B3 Mean counterfactual metrics for initial class STTC (n=148)

Target Class	Series	$Validity_{multi}$	L_0 sparsity	Noise stability	Margin
CD	Original	1.0000	0.9858	1.0000	0.6269
	Sparse	1.0000	0.8136	1.0000	0.6218
	Aligned Sparse	0.6923	0.7889	1.0000	0.5037
HYP	Original	1.0000	0.9808	1.0000	0.4057
	Sparse	1.0000	0.7485	1.0000	0.4013
	Aligned Sparse	0.0000	0.8803	0.9000	0.4431
MI	Original	1.0000	0.9826	1.0000	0.9404
	Sparse	1.0000	0.7479	1.0000	0.9343
	Aligned Sparse	1.0000	0.7284	1.0000	0.8790
NORM	Original	1.0000	0.9878	1.0000	0.6840
	Sparse	1.0000	0.9386	1.0000	0.6793
	Aligned Sparse	0.5733	0.9108	1.0000	0.4134

Table B4 Mean counterfactual metrics for initial class CD (n=119)

Target Class	Series	Validity _{multi}	L_0 sparsity	Noise stability	Margin
HYP	Original	1.0000	0.9789	1.0000	0.4410
	Sparse	1.0000	0.8936	1.0000	0.4406
	Aligned Sparse	0.0000	0.9493	1.0000	0.5537
MI	Original	1.0000	0.9775	1.0000	0.9505
	Sparse	1.0000	0.6954	1.0000	0.9427
	Aligned Sparse	0.9748	0.6762	1.0000	0.8987
NORM	Original	1.0000	0.9881	1.0000	0.6846
	Sparse	1.0000	0.9378	1.0000	0.6791
	Aligned Sparse	0.6970	0.8784	1.0000	0.4915
STTC	Original	1.0000	0.9845	1.0000	0.6100
	Sparse	1.0000	0.7854	1.0000	0.6079
	Aligned Sparse	0.5909	0.8513	0.9773	0.6195

Table B5 Mean counterfactual metrics for initial class MI (n=111)

Target Class	Series	Validity _{multi}	L_0 sparsity	Noise stability	Margin
CD	Original	1.0000	0.9856	1.0000	0.6215
	Sparse	1.0000	0.7767	1.0000	0.6183
	Aligned Sparse	0.8542	0.7775	1.0000	0.4828
HYP	Original	1.0000	0.9844	1.0000	0.4743
	Sparse	1.0000	0.7944	1.0000	0.4712
	Aligned Sparse	0.2500	0.7110	1.0000	0.2372
NORM	Original	1.0000	0.9889	1.0000	0.6922
	Sparse	1.0000	0.9374	1.0000	0.6872
	Aligned Sparse	0.7658	0.9033	0.9910	0.4937
STTC	Original	1.0000	0.9824	1.0000	0.5981
	Sparse	1.0000	0.7454	1.0000	0.5982
	Aligned Sparse	0.7184	0.8103	0.9854	0.6113

Table B6 Mean counterfactual metrics for initial class HYP (n=21)

Target Class	Series	Validity _{multi}	L_0 sparsity	Noise stability	Margin
CD	Original	1.0000	0.9826	1.0000	0.6257
	Sparse	1.0000	0.8186	1.0000	0.6287
	Aligned Sparse	0.7143	0.8076	1.0000	0.4438
MI	Original	1.0000	0.9804	1.0000	0.9710
	Sparse	1.0000	0.7297	1.0000	0.9616
	Aligned Sparse	1.0000	0.7324	1.0000	0.9142
NORM	Original	1.0000	0.9837	1.0000	0.7537
	Sparse	1.0000	0.9426	1.0000	0.7497
	Aligned Sparse	0.8095	0.8529	1.0000	0.5818
STTC	Original	1.0000	0.9837	1.0000	0.6210
	Sparse	1.0000	0.7516	1.0000	0.6205
	Aligned Sparse	0.7059	0.8011	1.0000	0.6001

Table B7 Mean counterfactual metrics for combined initial classes. (I)

Initial Class	Target Class	Series	Validity _{multi}	L_0 sparsity	Noise stability	Margin
CD+MI (n=74)	HYP	Original	1.0000	0.9812	1.0000	0.4300
		Sparse	1.0000	0.9080	1.0000	0.4304
		Aligned Sparse	0.2000	0.7334	0.8000	0.2321
	NORM	Original	1.0000	0.9868	1.0000	0.6992
		Sparse	1.0000	0.9423	1.0000	0.6948
		Aligned Sparse	0.6081	0.9375	0.9792	0.4581
	STTC	Original	1.0000	0.9818	1.0000	0.6040
		Sparse	1.0000	0.7394	1.0000	0.6011
		Aligned Sparse	0.6364	0.8195	0.9884	0.6336
CD+STTC (n=45)	HYP	Original	1.0000	0.9745	1.0000	0.4135
		Sparse	1.0000	0.8026	1.0000	0.4113
		Aligned Sparse	0.0000	0.7445	1.0000	0.3810
	MI	Original	1.0000	0.9824	1.0000	0.9424
		Sparse	1.0000	0.7473	1.0000	0.9347
		Aligned Sparse	0.9556	0.7090	1.0000	0.8996
	NORM	Original	1.0000	0.9873	1.0000	0.6883
		Sparse	1.0000	0.9490	1.0000	0.6852
		Aligned Sparse	0.6250	0.8732	1.0000	0.4941
HYP+STTC (n=39)	CD	Original	1.0000	0.9794	1.0000	0.5953
		Sparse	1.0000	0.8980	1.0000	0.5953
		Aligned Sparse	0.9231	0.7974	1.0000	0.5212
	MI	Original	1.0000	0.9746	1.0000	0.9513
		Sparse	1.0000	0.8007	1.0000	0.9429
		Aligned Sparse	1.0000	0.7802	1.0000	0.8942
	NORM	Original	1.0000	0.9822	1.0000	0.7464
		Sparse	1.0000	0.9411	1.0000	0.7445
		Aligned Sparse	0.8000	0.9234	1.0000	0.5173
MI+STTC (n=31)	CD	Original	1.0000	0.9843	1.0000	0.5927
		Sparse	1.0000	0.8364	1.0000	0.5929
		Aligned Sparse	0.6875	0.7992	1.0000	0.4547
	HYP	Original	1.0000	0.9851	1.0000	0.4410
		Sparse	1.0000	0.7961	1.0000	0.4414
		Aligned Sparse	0.0000	0.8477	1.0000	0.4405
	NORM	Original	1.0000	0.9880	1.0000	0.6948
		Sparse	1.0000	0.9510	1.0000	0.6884
		Aligned Sparse	0.6129	0.9172	0.9839	0.4295

Table B8 Mean counterfactual metrics for combined initial classes. (II)

Initial Class	Target Class	Series	Validity _{multi}	L_0 sparsity	Noise stability	Margin
CD+NORM (n=25)	HYP	Original	1.0000	0.9862	1.0000	0.4743
		Sparse	1.0000	0.7297	1.0000	0.4681
		Aligned Sparse	0.5000	0.9136	1.0000	0.3267
	MI	Original	1.0000	0.9814	1.0000	0.9799
		Sparse	1.0000	0.7301	1.0000	0.9710
		Aligned Sparse	1.0000	0.6587	1.0000	0.9362
	STTC	Original	1.0000	0.9860	1.0000	0.5683
		Sparse	1.0000	0.8183	1.0000	0.5718
		Aligned Sparse	0.5455	0.7837	1.0000	0.6298
CD+HYP (n=11)	MI	Original	1.0000	0.9780	1.0000	0.9559
		Sparse	1.0000	0.6796	1.0000	0.9446
		Aligned Sparse	1.0000	0.7296	1.0000	0.9022
	NORM	Original	1.0000	0.9871	1.0000	0.6395
		Sparse	1.0000	0.9414	1.0000	0.6381
		Aligned Sparse	0.8182	0.9621	1.0000	0.4175
	STTC	Original	1.0000	0.9840	1.0000	0.6257
		Sparse	1.0000	0.7686	1.0000	0.6271
		Aligned Sparse	0.6667	0.8813	1.0000	0.6440
HYP+MI (n=8)	CD	Original	1.0000	0.9830	1.0000	0.5671
		Sparse	1.0000	0.8913	1.0000	0.5725
		Aligned Sparse	1.0000	0.6807	1.0000	0.4500
	NORM	Original	1.0000	0.9873	1.0000	0.6380
		Sparse	1.0000	0.9301	1.0000	0.6402
		Aligned Sparse	0.7500	0.9688	1.0000	0.3957
	STTC	Original	1.0000	0.9754	1.0000	0.5286
		Sparse	1.0000	0.7159	1.0000	0.5265
		Aligned Sparse	0.6250	0.8909	0.8750	0.6000
NORM+STTC (n=4)	MI	Original	1.0000	0.9870	1.0000	0.9424
		Sparse	1.0000	0.7188	1.0000	0.9361
		Aligned Sparse	1.0000	0.7317	1.0000	0.8757

Appendix C Visualized instances of ECG counterfactuals from the proposed method

This appendix presents visual illustrations of counterfactual generation for ECG sample 1124. The ECG recording from test set (*Query*) classified as Conduction Disturbance (CD) appears in Figure C3, while Figure C4 shows the Hypertrophy (HYP) prototype (*Original*), which serves as an initial counterfactual by selecting the most representative sample from the HYP cluster, which is also similar to the given ECG sample. For this example, the counterfactual achieves perfect validity (*validity*: 1), successfully flipping the classifier’s prediction from CD to HYP. Subsequent overlays (Figures C5, C6) illustrate how the sparsified prototype (*Sparse*) highlights waveform adjustments required to effect the flip, and Figure C7 demonstrates adjustments with R-peak–based temporal alignment (*Alignment*). The combined alignment and sparsification step is shown in Figure C8 (*Sparse-Aligned*), and finally, Figure C9 overlays SHAP-derived attributions on the aligned, sparsified prototype to reveal which ECG features most strongly influence the model’s decision.

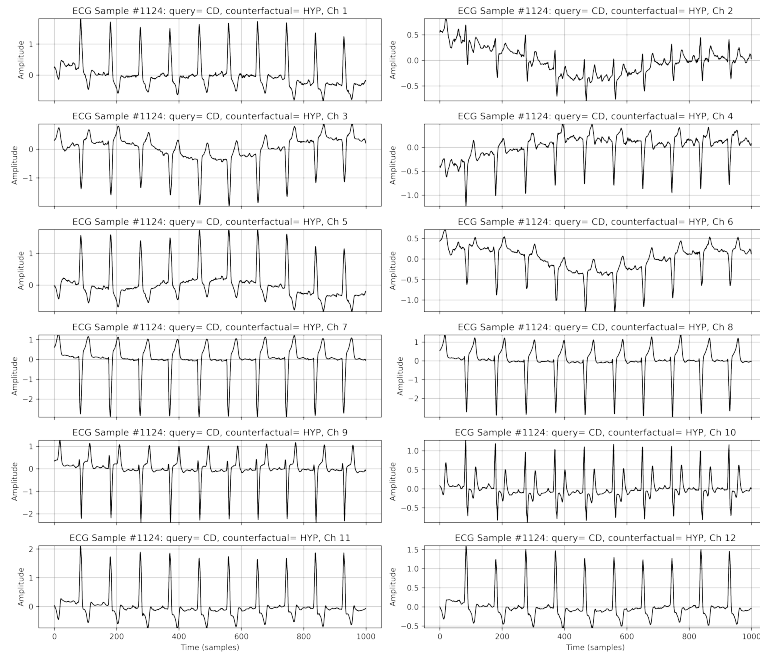


Fig. C3 ECG recording from test set (ID 1124) classified as Conduction Disturbance (CD); all 12 leads shown. (*Query*)

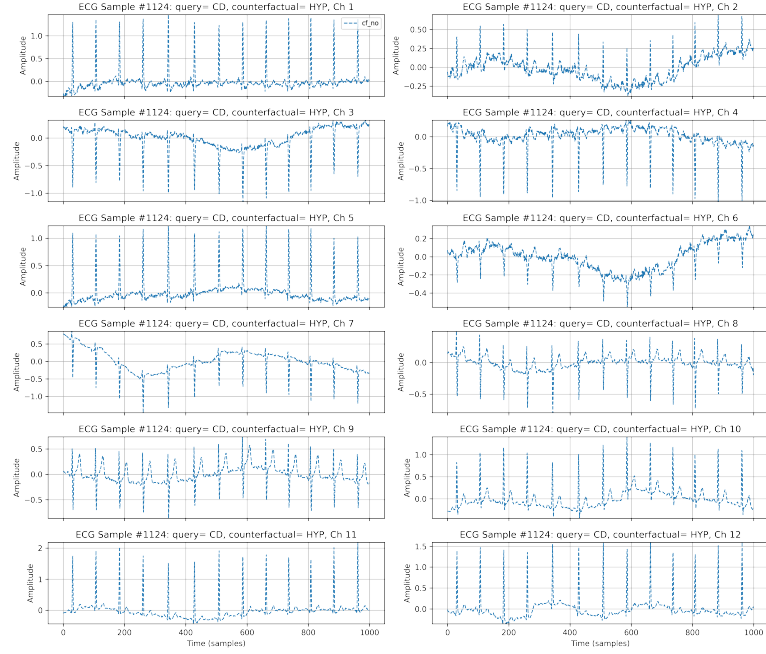


Fig. C4 Prototype counterfactual for Hypertrophy (HYP). (*Original*)

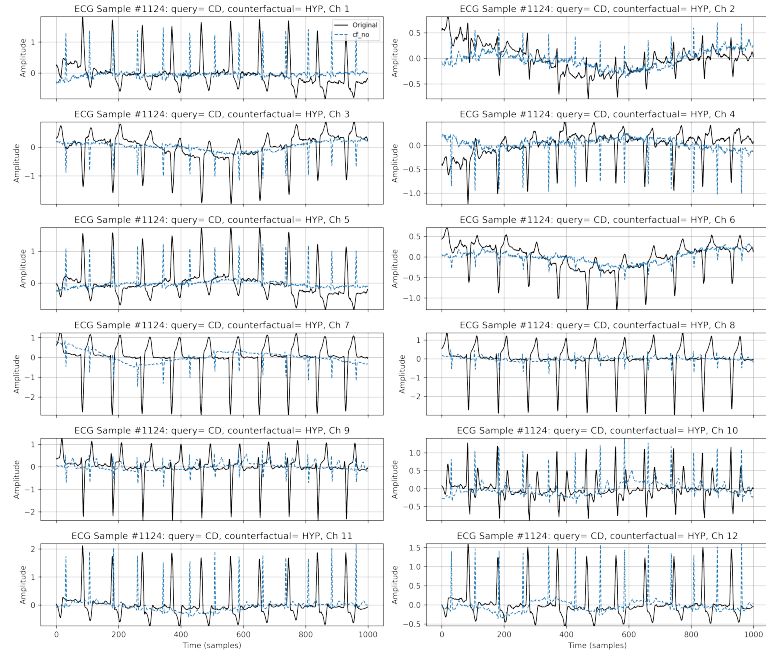


Fig. C5 Overlay of the ECG test sample and prototype. (*Query and Original*)

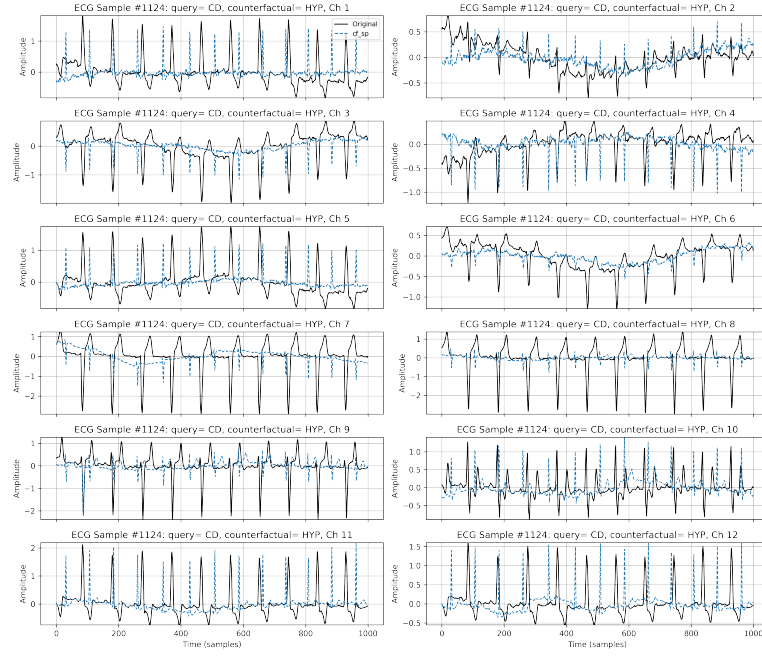


Fig. C6 Overlay of the ECG test sample and sparsified prototype, which highlights minimal feature changes. (*Query* and *Sparse*)

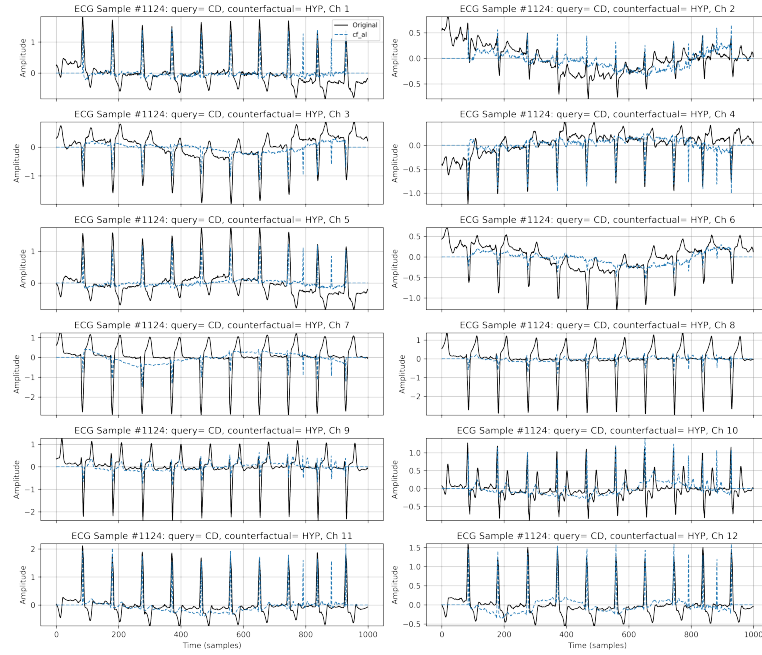


Fig. C7 Overlay of the ECG test sample and temporally aligned prototype. (*Query* and *Aligned*)

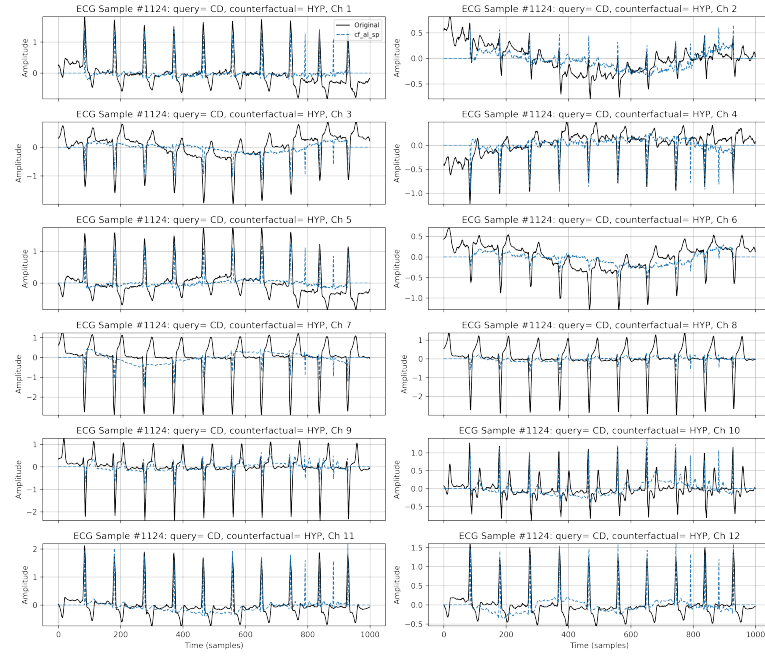


Fig. C8 Overlay of the ECG test sample and first aligned then sparsified prototype; minimal, physiologically plausible changes highlighted. (*Query* and *Sparse-Aligned*)

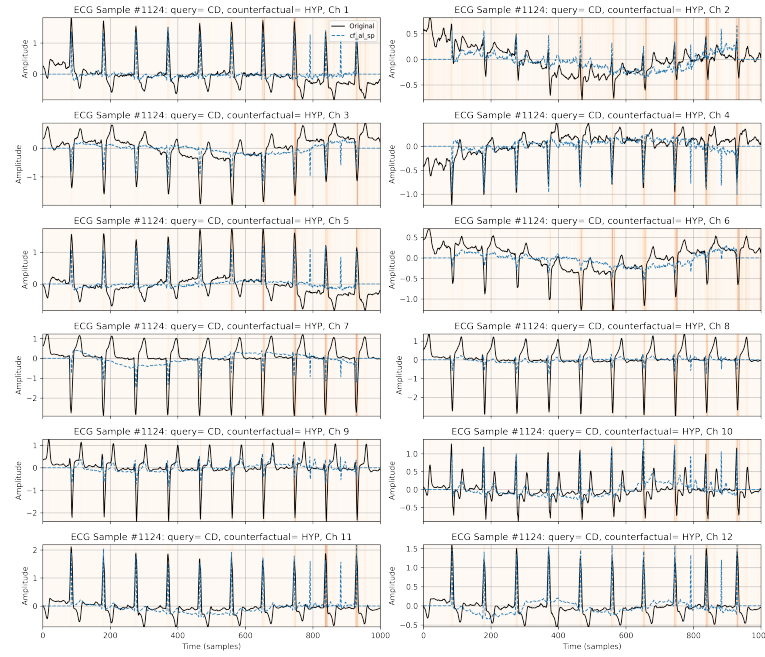


Fig. C9 Overlay of the ECG test sample and first aligned then sparsified prototype with SHAP heatmap superimposed. (*Query* and *Sparse-Aligned*)