Expert Systems With Applications 334 (2027) 134714



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Improved interpretability and feature attribution for hybrid–AI ECG models by controlling metadata leakage and feature interactions ⋆

Lennert Bontinck [a,*], Karel Fonteyn [a], Antônio H. Ribeiro [b,c], Tom Dhaene [a], Dirk Deschrijver [a]

[a] IDLab, Ghent University – imec, Technologiepark-Zwijnaarde 126, Ghent, 9052, Belgium
[b] Department of Information Technology, Uppsala University, Lägerhyddsvägen 1, Uppsala, 75237, Sweden
[c] Science for Life Laboratory, Husargatan 3, Uppsala, 75237, Sweden

ARTICLE INFO

Keywords:
Electrocardiography
Atrial fibrillation
Deep learning
Model interpretability
Metadata invariance
Feature attribution quantification
Out-of-distribution generalisation

ABSTRACT

Deep learning has shown strong potential for predicting atrial fibrillation (AFib) risk from electrocardiograms (ECGs). However, in hybrid-AI settings that combine an ECG latent representation with patient metadata (age and sex in this study), it can remain unclear to what extent predictions are driven by ECG latent features or by metadata. This ambiguity arises because the ECG latent representation may already encode metadata, causing attribution quantification methods to underestimate the attribution of explicitly provided metadata. A unified framework is proposed to improve interpretability by controlling metadata leakage, modelling ECG-metadata interactions explicitly, and enabling modality-aware attribution quantification. First, a Gradient-Reversed encoder (GRe) learns an ECG latent representation that is invariant to prespecified metadata, reducing metadata leakage. Second, a Multiplicative Feature-Interaction Dense (MFID) layer reintroduces metadata through gated multiplicative interactions with the ECG latent representation. Third, Feature-Interaction Gradient Interpretation (FIGI) combines ECG-level saliency visualisation using Grad-CAM++ with integrated gradients on a latent proxy model to quantify, per recording, the relative attribution of the ECG latent representation versus metadata in the final prediction. Across in-distribution and out-of-distribution evaluation, the proposed framework yields more clinically plausible attribution patterns and demonstrates that metadata can strongly influence AFib prediction, an effect that may be obscured by naive hybridisation. In the evaluated experiments, the proposed framework improved out-of-distribution (OOD) AFib prediction performance, although the magnitude of the benefit was task- and dataset-dependent.

1. Introduction

Deep Learning (DL) has shown considerable promise for analysing electrocardiogram (ECG) signals, achieving strong performance across a wide range of diagnostic and prognostic tasks (Fonteyn et al., 2025; Lin et al., 2025; Strodthoff et al., 2024). By learning directly from raw ECGs, such models can capture subtle ECG signal characteristics that may be missed by conventional feature engineering. Atrial Fibrillation (AFib) is a particularly compelling application: it is among the most common Cardiovascular Diseases (CVDs), is associated with substantial morbidity, and earlier identification provides clear clinical benefit (Hindricks et al., 2020). Recent studies have shown that DL models can identify AFib-positive patients from ECGs acquired during Normal Sinus Rhythm (NSR) after a prior AFib episode (Attia et al., 2019; Yuan et al., 2023). In some settings, they can also predict future AFib from ECGs recorded before a patient's first documented episode (Bontinck et al., 2025; Tsiartas et al., 2025). Across these tasks, performance can match or surpass established clinical risk scores such as CHARGE-AF and the FHS score (Brant et al., 2025; Tsiartas et al., 2025).

Despite these advances, conventional risk scores remain popular because their structure is transparent and easy to interpret. They combine a small set of well-understood variables, each entering the model with an explicit, fixed contribution to the final risk prediction. These variables are often patient metadata (e.g., age and sex) and are termed Clinically Interpretable Features (CLFs) throughout this paper. In contrast, DL models are often treated as black boxes, and the role of individual

⋆ Funding: This work was conducted as part of the VASCULAI project, supported by the European Union through the Interreg France–Wallonie–Vlaanderen (FWVL) program (ref. 0100224) and the province Oost-Vlaanderen. The research was further supported by the Flemish Government via the AI Research Program.

* Corresponding author.
E-mail addresses: lennert.bontinck@ugent.be (L. Bontinck), karel.fonteyn@ugent.be (K. Fonteyn), antonio.horta.ribeiro@it.uu.se (A.H. Ribeiro), tom.dhaene@ugent.be (T. Dhaene), dirk.deschrijver@ugent.be (D. Deschrijver).

https://doi.org/10.1016/j.eswa.2026.134714
Received 27 February 2026; Received in revised form 29 September 2026; Accepted 10 October 2026
Available online 15 October 2026
0957-4174/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

ECG latent features can be difficult to assess. Hybrid-AI approaches attempt to bridge this gap by combining an ECG latent representation with explicit CLFs, frequently improving predictive performance (Bontinck et al., 2026; Brant et al., 2025; Rawshani et al., 2025). However, naive hybridisation (e.g., simple concatenation) can create an interpretability trap, particularly when CLFs are already reflected in the ECG latent representation.

A central difficulty is that key CLFs are often highly predictable from the ECG. For example, age and sex can be inferred with high accuracy from standard 12-lead recordings using DL models (Barros et al., 2024). This implies that ECG encoders trained for AFib classification may already encode these CLFs implicitly in the ECG latent representation. Such *CLF leakage* confounds interpretation in two ways. First, adding these CLFs as explicit inputs may yield only modest gains in predictive performance, even when they are clinically informative. Second, attribution analyses may assign little importance to the explicit CLFs inputs, even if the model is effectively relying on the same information via its latent representation.

This paper argues that meaningful interpretation of CLFs in hybrid-AI requires controlling *where* CLFs enter the model and how they interact with the ECG latent features. Accordingly, a unified framework with three key contributions is proposed. First, a Gradient-Reversed encoder (GRe) learns an ECG latent representation that is explicitly invariant to prespecified CLFs (age and sex in this study). Second, Multiplicative Feature-Interaction Dense (MFID) reintroduces these CLFs through gated multiplicative interactions with ECG latent features, providing an explicit mechanism for modelling how CLFs modulate ECG-derived evidence. Third, Feature-Interaction Gradient Interpretation (FIGI) combines ECG-level saliency visualisation using Grad-CAM++ with integrated-gradients-based attribution on a latent proxy model to quantify, for each model prediction, the relative attribution of CLFs versus ECG latent features (Chattopadhay et al., 2018; Sundararajan et al., 2017). The proposed framework is not tied to a single classifier architecture; rather, it provides a controlled way to test how hybrid-AI classifiers use explicitly provided CLFs relative to CLF information already present in the ECG latent representation.

Through experiments on a large-scale in-distribution data set and an external Out-of-Distribution (OOD) evaluation data set, it is shown that suppressing CLF leakage is necessary for reliable attribution quantification. Naive hybridisation based on simple concatenation and conventional dense layers can obscure the contribution of explicit CLFs, whereas the proposed framework provides clinically plausible ECG saliency patterns and CLF attributions. Finally, the proposed framework is shown to improve OOD performance in selected task and model settings, particularly for the models most affected by CLF leakage.

2. Related work

This section summarises prior work relevant to this manuscript and identifies limitations that motivate the proposed GRe–MFID–FIGI framework. Specifically, prior work is reviewed on: (i) ECG-based AFib detection and risk prediction, (ii) concept-suppression methods for learned representations, (iii) neural architectures for explicit multiplicative interactions, and (iv) post-hoc interpretability techniques for ECG models.

2.1. AFib risk prediction

Traditional AFib risk prediction is commonly based on multivariable models built from CLFs such as age, sex, and comorbidities (Hindricks et al., 2020). Widely used examples include CHARGE-AF by Alonso et al. (2013), the FHS score by Schnabel et al. (2009) and $HARMS_2$-AF score by Segan et al. (2023). These models are attractive in clinical practice because they are simple and transparent, but they are restricted to predefined CLFs and cannot directly incorporate information contained in ECG signals.

In contrast, modern DL approaches demonstrate that raw 12-lead ECGs contain rich diagnostic and prognostic information. Strong performance has been reported for prevalent AFib detection and for predicting future AFib from earlier non-AFib recordings, suggesting that ECG-derived biomarkers may be present before AFib is documented in the available record sequence (Attia et al., 2019; Bontinck et al., 2025; Fonteyn et al., 2025; Yuan et al., 2023). Several studies report performance comparable to, or exceeding, standard clinical approaches and risk scores across cohorts and task formulations (Brant et al., 2025; Tsiartas et al., 2025).

A key open question is how much of such predictive performance reflects pure ECG signal characteristics versus CLFs that are implicitly encoded in an ECG latent representation. Age and sex can be inferred from 12-lead ECGs with high accuracy using DL models similar to those used for AFib classification (Barros et al., 2024). In addition, the discrepancy between a patient's age and DL-estimated *ECG age* has been reported to associate with AFib and broader CVDs (Cho et al., 2024; Lima et al., 2021b; Ribeiro & Ribeiro, 2024). Consequently, hybrid-AI models that concatenate an ECG latent representation with explicit CLFs can be redundant because the same CLF information may be present both explicitly and implicitly. While such simple hybridisation can improve predictive performance as shown by Brant et al. (2025), it offers limited insight into whether explicit CLFs provide additional information beyond what is already encoded in the ECG latent representation. As a result, the importance of CLFs for the final prediction can remain unclear. Moreover, simple concatenation does not explicitly model how CLFs modulate ECG latent features.

2.2. Suppressing concepts in learned representations

Methods that remove specified information from DL representations are often discussed in the fairness and bias-mitigation literature. In that setting, *sensitive* or *protected* attributes (e.g., race or sex) are treated as variables that should not be exploitable by the model, and the goal is to learn representations that remain predictive of the target while being invariant to those attributes (Parraga et al., 2025).

A prominent family of approaches relies on adversarial learning. Gradient Reversal (GR), introduced for unsupervised domain adaptation, enforces invariance by training an auxiliary predictor for the attribute while reversing its gradients at the encoder level (Ganin & Lempitsky, 2015). Related directions include adversarial disentanglement, invariant risk minimisation, and explicit concept erasure (Parraga et al., 2025; Yang et al., 2024).

In biomedical signal analysis, an analogous phenomenon arises naturally. CLFs such as age and sex shape baseline ECG signal characteristics and can therefore be encoded implicitly by DL models, even when not provided as explicit inputs or targets. Existing work has used domain adaptation and debiasing techniques in ECG settings to suppress acquisition- or dataset-related information when biases are present (Dong et al., 2025; Gliner et al., 2023). However, for AFib risk prediction, age and sex are not nuisance variables to be discarded, but established risk factors that should ultimately inform the final decision. This motivates a two-stage design: (i) enforce invariance to prespecified CLFs in the ECG latent representation, and (ii) reintroduce CLFs downstream through a controlled and interpretable pathway.

2.3. Explicit multiplicative interactions in neural networks

Multiplicative interactions have been studied as mechanisms for feature modulation, conditional computation, and relational reasoning. Examples include multiplicative MLPs, multiplicative recurrent units, and various gating mechanisms (Jayakumar et al., 2020; Schmitt, 2002; Wu et al., 2016). These works show that an explicit multiplicative structure can increase expressiveness beyond purely additive transformations, particularly when one set of inputs should modulate the effect of another.

Standard dense layers combine inputs additively before applying non-linearities. While universal approximation holds in principle, multiplicative dependencies must then be learned implicitly and are typically approximated locally in the joint input space. This local approximation can demand additional model capacity and dense in-distribution coverage, and it may be brittle under distribution shift (Schmitt, 2002).

In hybrid-AI ECG modelling, this point is practically important: CLFs such as age and sex often act as context variables that modulate how physiological patterns should be interpreted. Despite this, most hybrid pipelines rely on concatenation followed by standard dense layers, leaving clinically meaningful CLF–ECG interactions implicit.

2.4. Post-hoc interpretability for ECG models

Post-hoc interpretability methods aim to probe black-box DL models by identifying which parts of the input support a given prediction. For hybrid-AI ECG models, it is useful to distinguish between (i) *localisation* methods that highlight relevant regions in the ECG and (ii) *attribution quantification* methods that quantify the attribution of distinct input variables (the ECG versus explicit CLFs).

2.4.1. Signal localisation

Raw gradient saliency maps can be noisy for ECGs because individual *time samples* (i.e., single time points) have limited meaning in isolation (Simonyan et al., 2013). SmoothGrad addresses this by averaging input gradients over multiple perturbed versions of the same input (typically generated by adding random noise), which tends to suppress high-frequency artefacts and stabilise saliency patterns (Smilkov et al., 2017).

Grad-CAM produces saliency maps from deep activation maps (Selvaraju et al., 2020). Grad-CAM++ extends Grad-CAM with a refined weighting scheme that can better handle multiple occurrences of relevant patterns and improves localisation when the target depends on several distinct regions (Chattopadhay et al., 2018). In ECG modelling, the 12 leads are typically treated as input channels. As such, Grad-CAM-style methods provide primarily *temporal* localisation along the time axis of the selected convolutional feature maps (Bontinck et al., 2024).

2.4.2. Feature attribution

For scalar CLFs, attribution methods that are explicitly motivated by desirable theoretical properties (e.g., completeness and sensitivity) are often preferred because they aim to produce quantitatively meaningful attributions. Integrated Gradients (IG) attributes an output difference to inputs by integrating gradients along a path from a baseline input to the observed input (Sundararajan et al., 2017). In practice, this path integral is approximated numerically. The input is evaluated at a finite number of equally spaced interpolation points between the baseline and the observed input, and the gradients at these points are summed to approximate the integral. These interpolation points are the *integration steps*; using more steps gives a finer approximation, but increases computation. When implemented with a fixed baseline, IG satisfies desirable properties such as completeness (the attributions sum to the output difference relative to a baseline) and sensitivity. Its behaviour depends strongly on the baseline choice, which is straightforward for scalar CLFs but difficult for raw ECGs because physiologically neutral reference waveforms are not well defined.

Deep Shapley Additive Explanations (DeepSHAP) combines DeepLIFT-style rules with SHAP's additive attribution framework and can be more stable when an appropriate background distribution is available (Chen et al., 2022). However, results depend on the chosen background set and the network structure.

2.4.3. Hybrid interpretability challenges

Applying attribution methods directly to hybrid models that combine scalar CLFs with the high-dimensional ECG presents two recurring difficulties. First, the ECG dimensionality (e.g., 5000×12 time samples) can numerically dominate aggregated attribution scores, masking the attribution of scalar CLFs. Second, if CLFs are implicitly encoded in an ECG latent representation, attribution can misleadingly assign importance to the ECG modality for effects that are in fact CLF-driven. These limitations motivate interpretability frameworks that decouple modalities and explicitly control CLF leakage before quantifying CLF attributions.

2.5. Summary and identified gaps

In summary, prior work has independently addressed (i) ECG-based AFib prediction, (ii) suppression of specified information in learned representations, (iii) multiplicative feature interactions, and (iv) post-hoc interpretability for hybrid-AI ECG models. However, these lines of work remain largely disconnected in existing literature for AFib prediction. In particular, hybrid models typically do not control *where* CLFs enter the model. Clinically meaningful modulation of ECG latent features by CLFs is often left implicit. In addition, attribution methods are frequently applied in settings where the ECG latent representation already encodes CLFs, which undermines any apparent *ECG-only* explanation. This gap motivates a unified framework that (i) enforces CLF invariance in the ECG representation, (ii) reintroduces CLFs through an explicit interaction mechanism, and (iii) supports modality-aware interpretation of ECG versus CLF attributions in clinically relevant settings.

3. Proposed architecture

This section describes a three-stage hybrid-AI framework for ECG classification in which an ECG latent representation and explicit CLFs interact in a controlled and interpretable manner. First, a Gradient-Reversed encoder (GRe) enforces CLF invariance in the ECG latent representation, preventing unintended CLF leakage. Second, a Multiplicative Feature-Interaction Dense (MFID) layer reintroduces CLFs through gated multiplicative interactions with ECG latent features, producing the final prediction while making the interaction mechanism explicit. Third, Feature-Interaction Gradient Interpretation (FIGI) supports post-hoc interpretation by combining signal localisation and modality-level attribution.

3.1. GRe: Gradient-reversed encoder

As discussed in Section 2.2, Gradient Reversal (GR) was originally introduced for unsupervised domain adaptation (Ganin & Lempitsky, 2015). In the present setting, GRe uses the same gradient-reversal mechanism to train an ECG encoder that remains predictive of the primary downstream task while suppressing information about prespecified CLFs.

Fig. 1 illustrates the approach. A shared encoder maps the raw ECG to an ECG latent representation that feeds multiple output branches. One branch predicts the clinical target (i.e., AFib), while auxiliary branches predict the prespecified CLFs (i.e., age and sex). A GR layer is inserted between the encoder and each CLF head. During training, all losses are minimised, but gradients flowing back from the CLF heads are multiplied by a negative factor in the GR component. As a result, the CLF heads learn to predict the CLFs, while the encoder is encouraged to remove information that supports CLF prediction. This differs from simply maximising the CLF loss, which can be satisfied by trivial label inversions rather than true invariance.

The extent of CLF removal is assessed by training probe models to reconstruct age and sex from frozen ECG latent representations (Section 4.3).

3.1.1. Technical implementation

The GR layer acts as an identity map during the forward pass and multiplies the incoming gradient by $-\lambda$ during the backward pass. A λ

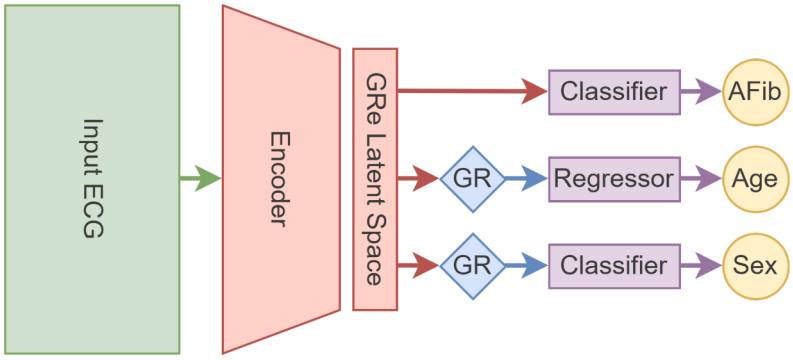


Fig. 1. Schematic overview of GRe training. A shared ECG encoder feeds a primary target head (i.e., AFib) and auxiliary CLF heads (i.e., age and sex). A GR layer reverses gradients from the CLF heads (scaled by λ), encouraging the encoder to remove CLF information from the latent representation. Separate gradient tapes enable a weighted combination of branch gradients for the encoder update.

scheduler is employed to avoid destabilising optimisation early in training. Let t denote the optimisation step, b the burn-in duration, w the warm-up duration, and $\lambda_{\max}$ the maximum reversal strength. The schedule used in this study is provided in (1).

$$\lambda(t) = \lambda_{\max} \cdot \mathrm{clip}\left(\frac{t-b}{\max(1,w)}, 0, 1 \right) \tag{1}$$

In this equation, $\mathrm{clip}(\cdot, 0, 1)$ truncates values outside $[0, 1]$. Thus, $\lambda(t) = 0$ throughout burn-in and then increases linearly until $\lambda_{\max}$, which was set to 1. In this manuscript, burn-in and warm-up correspond approximately to 10 and 60 training epochs, respectively.

Separate gradient tapes are maintained for each output branch, enabling explicit control over the contribution of each loss to the shared encoder update. This permits a weighted combination of branch gradients, which is useful when strong CLF penalisation would otherwise reduce the primary task performance.

3.2. MFID: Multiplicative feature-interaction dense

With a GRe-trained ECG latent representation that emphasises ECG signal characteristics while suppressing prespecified CLFs, the downstream classifier must reintroduce CLFs explicitly. Standard hybrid-AI approaches typically concatenate CLFs with an ECG latent representation and rely on dense layers to approximate interactions implicitly. In contrast, MFID is designed to represent modulatory effects explicitly. For example, a latent feature can be a strong marker for AFib in younger patients but less informative in older patients. Such behaviour is naturally multiplicative: the influence of a latent feature depends on the CLF value.

MFID therefore combines a conventional dense path with a gated multiplicative path. The gating enables the model to suppress unnecessary interactions, providing a fallback to additive dense behaviour when a multiplicative relation is not present. Section 4.4 exemplifies a setting in which the evaluated standard Dense approaches generalise poorly and illustrates how explicitly modelling multiplicative relations with MFID can improve performance under distribution shift.

3.2.1. Technical implementation

The MFID operation is defined in (2) and illustrated in Fig. 2. Let $\mathbf{Z} = (z_1, \ldots, z_L) \in \mathbb{R}^L$ denote the ECG latent representation produced by the encoder, and let $\mathbf{C} = (c_1, \ldots, c_I) \in \mathbb{R}^I$ denote the vector of CLFs. The gating activation is denoted by $\phi_{\mathrm{gate}}(\cdot)$ and the output activation by $\phi_{\mathrm{out}}(\cdot)$. All weights ($w$) and biases ($b$) are learnable parameters. The layer output aggregates a standard dense path (h_{dense}) and the sum over all pairwise gated interaction terms (h_{int}).

$$\begin{aligned} h_{\mathrm{dense}} &= \sum_{l=1}^{L} z_l \, w_l^{\mathrm{sd}} + b^{\mathrm{sd}}, \\ g_{i,l} &= \phi_{\mathrm{gate}}\left(z_l \, w_{i,l}^{\mathrm{gate}} + b_{i,l}^{\mathrm{gate}} \right), \\ u_{i,l} &= c_i \, w_{i,l}^{\mathrm{clf}} + b_{i,l}^{\mathrm{clf}}, \\ h_{\mathrm{int}} &= \sum_{i=1}^{I} \sum_{l=1}^{L} g_{i,l} \, u_{i,l}, \\ y &= \phi_{\mathrm{out}}\left(h_{\mathrm{dense}} + h_{\mathrm{int}} \right). \end{aligned} \tag{2}$$

For the gating activation ϕ_{gate}, LeakyReLU is used to allow negative values while making it easy to learn near-zero gates that effectively disable specific interactions. The output activation ϕ_{out} is task dependent but is also set to LeakyReLU in the reported experiments. To preserve the possibility of a conventional additive fall-back, the latent input $\mathbf{z}$ can be constructed to include both the previous layer output and a transformed representation of the CLFs.

3.3. FIGI: Feature-interaction gradient interpretation

FIGI provides two complementary views of model behaviour: (i) localisation of discriminative regions in the raw ECG waveform and (ii) quantification of the relative attribution of CLFs versus the ECG latent representation.

For localisation, Grad-CAM++ (Chattopadhay et al., 2018) is applied to the end-to-end model. As discussed in Section 2.4, when the encoder treats the 12 leads as input channels, Grad-CAM++ primarily yields temporal regions of interest, because lead-specific information is reduced by the channel-wise aggregation inherent to the method. Representative Grad-CAM++ maps are shown in Fig. 5.

For modality-level attributions, Integrated Gradients (IG) is used because it satisfies completeness (attributions sum to the output difference relative to a baseline) and can be applied to heterogeneous inputs (Sundararajan et al., 2017). As described in Section 2.4.2, the integration steps used by IG are the interpolation points used to approximate the path integral between a baseline input and the observed input. However, applying IG directly to the raw ECG together with scalar CLFs is ill-conditioned: the ECG dimensionality (i.e., 5000×12) numerically dominates aggregated attribution scores, and defining a physiologically meaningful baseline waveform is non-trivial. FIGI therefore performs IG on a latent proxy model.

3.3.1. Technical implementation

To quantify CLFs attributions, the trained model is split at the point where the ECG latent representation is combined with the scalar CLFs. The encoder first maps the raw ECG to a latent vector $\mathbf{z}$. The remaining

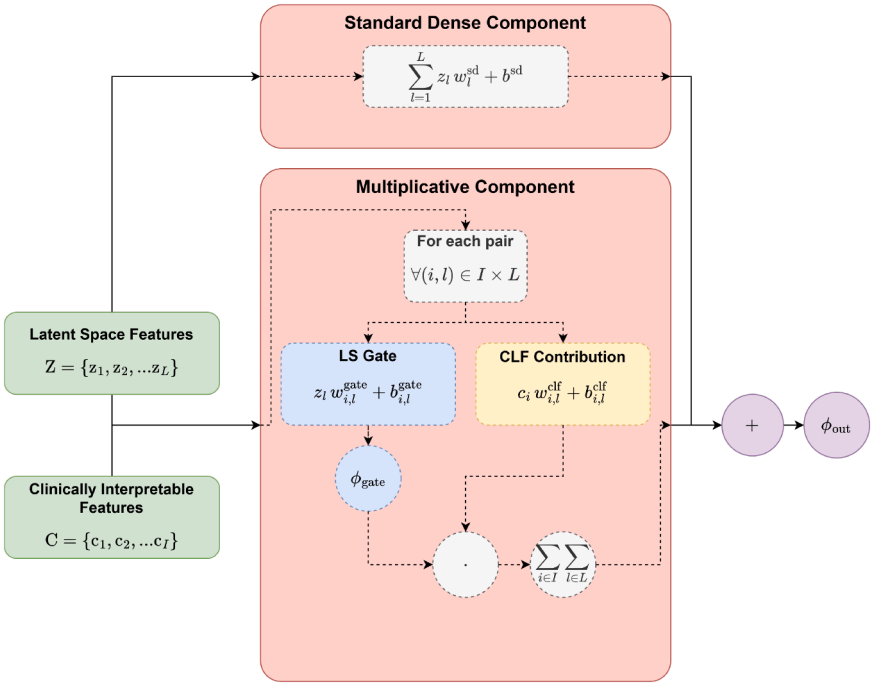


Fig. 2. Schematic overview of MFID layer. MFID combines two parallel branches: (i) a standard dense path that takes only **Z** as input, and (ii) a gated multiplicative path that takes **Z** and **C** as input and models pairwise interactions between CLFs and latent features. Gates close to zero suppress unnecessary interactions, enabling a fall-back towards additive behaviour.

trained classification head then combines **z** with age and sex to produce the final prediction. The Latent Proxy Model ($F_{\mathcal{L}}$) is this trained classification head evaluated directly on the pre-computed latent vector and scalar CLFs: $\hat{y} = F_{\mathcal{L}}(\mathbf{z}, c_{\mathrm{age}}, c_{\mathrm{sex}})$

No additional model is trained for this step: $F_{\mathcal{L}}$ reuses the exact trained weights of the original classification head. For a fixed latent vector **z** produced by the encoder, it therefore represents the same downstream decision function as the original model, but with the raw ECG replaced by its latent representation. This makes the attribution problem better conditioned, because baselines can be defined for the latent representation and for the scalar CLFs, while the very high-dimensional raw ECG no longer dominates the summed attribution scores. Raw-signal localisation is handled separately by Grad-CAM++, while the latent proxy is used only for modality-level attribution. IG is computed on $F_{\mathcal{L}}$ with $m = 50$ integration steps, meaning that the integral is approximated using 50 interpolation points between each baseline and the corresponding observed input. The following baselines are used:

- **Latent baseline ($\mathbf{z}'$):** the latent-space centroid, i.e., the mean latent vector $\mathbf{z}' = \frac{1}{N}\sum_{n=1}^{N} \mathbf{z}^{(n)}$ averaged over the N training recordings.
- **Age baseline:** population mean age.
- **Sex baseline:** counterfactual baseline $1 - c_{\mathrm{sex}}$ (binary sex label setting).

The importance of the ECG modality is obtained by summing absolute attributions across all dimensions of **z**, while CLF importances are the absolute attributions of the corresponding scalar inputs.

4. Evaluation and results

The proposed hybrid-AI framework is evaluated through experiments that validate the individual components (GRe, MFID, and FIGI) and assess their combined effect on downstream AFib classification. Four objectives are addressed: (i) verification that GRe suppresses prespecified CLFs (age and sex) in the ECG latent representation (Section 4.3); (ii) a synthetic proof-of-concept demonstrating that MFID captures multiplicative relations that standard dense layers may approximate poorly, particularly under distribution shift (Section 4.4); (iii) downstream AFib classification performance on in-distribution CODE-15% and OOD PTB-XL, including age- and sex-stratified analysis of the Prediction task; and (iv) model-decision analysis with FIGI to localise important ECG regions and to quantify the relative attributions of the ECG latent representation versus CLFs for model predictions (Section 4.6). The following subsections describe the data sets, experimental protocol, and results.

4.1. Available data sets

Two large-scale 12-lead ECG data sets are used to evaluate the proposed framework. The CODE-15% data set by Ribeiro et al. (2020) serves as the primary source for model training, validation, and in-distribution testing (Lima et al., 2021a; Ribeiro et al., 2021). The PTB-XL data set by Wagner et al. (2020) is employed exclusively for OOD evaluation to assess generalisation across populations and acquisition conditions (Goldberger et al., 2000; Wagner et al., 2022).

4.1.1. CODE-15%

The CODE-15% data set is a stratified 15% subset of the Clinical Outcomes in Digital Electrocardiology (CODE) registry and contains 345,779 12-lead ECG recordings acquired between 2010 and 2016 through the Telehealth Network of Minas Gerais in Brazil. These recordings originate from 233,770 unique patients. The cohort has a mean patient age of 53 ± 20 years, and 40% of the patients are male.

Original signals were recorded at a sampling frequency of 400 Hz and consist of both 10-s recordings and zero-padded 7-s segments. For compatibility with external evaluation data, zero-padding was removed prior to preprocessing, after which all signals were resampled to 500 Hz. Seven-second recordings were subsequently re-padded to obtain a uniform input shape of $12 \times 5{,}000$ samples. As device-specific acquisition information is not available for CODE-15%, a global normalisation strategy was applied across all recordings.

For model development, the data set was split into training, validation, and test partitions using a 10-fold, per-patient splitting scheme to avoid patient overlap. Additionally, folds were stratified by diagnostic labels to maintain class balance. Seven folds were used for training, one for validation, and two for testing.

4.1.2. PTB-XL

The PTB-XL data set is used to evaluate robustness under distribution shift and represents a distinct population and acquisition setting. It consists of 21,837 standardised 10-s 12-lead ECG recordings collected from 18,885 patients between 1989 and 1996. All signals are sampled at 500 Hz, yielding an input shape of $12 \times 5{,}000$. The patient population has a mean age of 63 ± 32 years, with 52% of patients being male.

In contrast to CODE-15%, PTB-XL includes detailed information on ECG acquisition devices, enabling device-specific normalisation. To ensure a strict OOD evaluation, no PTB-XL recordings were used during training, validation, or threshold optimisation. Instead, all predefined folds of the data set were reserved exclusively for testing, providing a comprehensive assessment of generalisation across both CLF and signal acquisition shifts.

4.2. Experimental setup and evaluation metrics

The experimental pipeline is organised into two sequential stages. In the first stage, an ECG encoder is trained either conventionally or with the proposed GRe mechanism to learn an ECG latent representation. In the second stage, the encoder is frozen and a downstream classifier (Dense or MFID) is trained, explicitly reintroducing the CLFs.

4.2.1. Encoder architectures and stage 1 training

To investigate the influence of model capacity on CLF leakage and downstream behaviour, two representative encoder architectures with markedly different representational capacities are considered: a compact encoder made specifically for ECG signals (ECGencode) and a commonly used high-capacity architecture (Attia).

- **ECGencode:** A compact convolutional architecture optimised for computational efficiency and deployment proposed by Bontinck et al. (2024). When configured according to the ECGencode Model 1 setup, it produces an ECG latent representation of dimensionality 640.
- **Attia:** A higher-capacity architecture inspired by a ResNet-8 backbone, following the design proposed by Attia et al. (2019) and commonly adopted in subsequent work (Bontinck et al., 2025; Raghunath et al., 2021). This model yields a substantially larger ECG latent representation of dimensionality 15,360.

Both encoders are trained on a "related" AFib classification task. At the patient level, all ECGs belonging to individuals with any documented AFib diagnosis are labelled positive, while recordings from patients without AFib remain negative. This formulation, referred to as the *Related* task, increases the effective number of positive samples. It includes three positive-recording groups: ECGs labelled as AFib (Detection), non-AFib-labelled ECGs recorded before the first documented AFib-positive ECG (Prediction), and non-AFib-labelled ECGs recorded afterwards (NSR). All task definitions used for the experiments are summarised in Table 1.

The Prediction task is defined by the temporal ordering of each patient's recordings: a non-AFib-labelled ECG is a positive Prediction sample when it precedes that patient's first documented AFib-positive ECG. In the CODE-15% test set, these prediction ECGs were acquired a mean of 600.9 ± 535.5 days before the first AFib-positive ECG (median: 459 days). In PTB-XL, the corresponding ECGs preceded the first AFib-positive ECG by 117.6 ± 228.6 days on average (median: 14.1 days).

For each encoder, three training configurations are evaluated:

1. **Regular:** Conventional end-to-end training for AFib classification without explicit constraints on the latent representation.
2. **GRe:** Training with the proposed GRe technique, where a loss weight of 4 is assigned to the primary AFib task and a weight of 1 to each auxiliary CLF branch (age and sex). This configuration prioritises task performance while encouraging CLF invariance.
3. **Balanced GRe:** Training with equal loss weights across all branches, enforcing stronger CLF invariance at the potential expense of predictive performance.

Across all Stage 1 experiments, classification heads consist of three dense layers with 128, 64, and 1 output unit. For GRe-based variants, the auxiliary CLF heads are preceded by a GR layer following the setup shown in Fig. 1.

4.2.2. Downstream classification and stage 2 training

In the second stage, encoder weights are frozen and a final classifier is trained for the downstream task. At this stage, CLFs are explicitly reintroduced. All CLF inputs are first normalised and then passed through a projection block comprising a dense layer with a single output (allowing the model to learn a scalar offset), followed by a dense layer with eight outputs to obtain a low-dimensional learned representation.

Two classifier heads are compared:

- **Dense baseline:** A standard classifier consisting of three dense layers (128, 64, and 1 unit). The projected CLF representations (eight features per CLF) are concatenated with the latent representation at each intermediate layer, starting at the encoder output.
- **MFID:** The proposed MFID-based classifier, which is identical to the baseline with the exception of the first dense layer being replaced with an MFID layer with 128 units. As in the baseline, the projected CLFs are concatenated with the latent space at each intermediate representation, while the normalised (non-projected) CLFs are used to drive the MFID gating mechanism.

4.2.3. Evaluation metrics

To quantify the presence of CLF concepts in the learned ECG latent representations (Section 4.3), auxiliary probe models are trained to predict age and sex from frozen latent vectors. Performance for sex prediction is reported using binary accuracy and Area Under the Receiver Operating Characteristic Curve (AUROC), while age prediction performance is assessed using the Mean Absolute Error (MAE). The Mean Squared Error (MSE) is used to evaluate the performance of MFID in the synthetic experiment (Section 4.4).

For the downstream AFib classification experiments (Section 4.5), performance is evaluated using AUROC and F1-score. To reflect a clinically realistic deployment scenario, the decision threshold used to compute the F1-score is optimised exclusively on the CODE-15% validation set and then applied unchanged to the PTB-XL test set. For the subgroup analysis of the Prediction task, AUROC is reported per age and sex subgroup, together with the number of positive samples in each subgroup.

4.3. Presence of CLF concepts in ECG latent representations

The effectiveness of GRe in suppressing CLF information is evaluated by measuring how accurately auxiliary models can reconstruct age and

Table 1
Definition of the evaluated AFib classification tasks. For all tasks, ECGs from AFib-negative patients are used as negative samples. Positive samples are obtained by task-specific filtering of ECGs from AFib-positive patients.

Task	Positive-sample criteria
Detection	ECGs labelled as AFib
Prediction	Non-AFib-labelled ECGs recorded before the first documented AFib-positive ECG
NSR	Non-AFib-labelled ECGs recorded after the first documented AFib-positive ECG
Related	All ECGs (i.e., the union of Detection, Prediction, and NSR)

Table 2
Quantitative evaluation of CLF leakage in latent representations. Auxiliary probe models are trained to predict age and sex from frozen latent vectors produced by the encoders described in Section 4.2 (Regular, GRe, and B-GRe for *Balanced* GRe). Rows are ordered from Regular to GRe and B-GRe to reflect increasing intended CLF suppression. For this analysis, lower sex prediction performance and higher age MAE indicate stronger suppression of CLF information. The constant-predictor reference reports majority-class sex prediction and train-set mean-age prediction applied to the test set.

Encoder	Model Variant	Sex Accuracy	Sex AUROC	Age MAE
ECGencode	Regular	0.6318	0.7196	14.48
	GRe	0.4105	0.5888	15.04
	B-GRe	0.4037	0.5637	15.71
Attia	Regular	0.7360	0.8053	11.55
	GRe	0.4834	0.5725	16.02
	B-GRe	0.4215	0.5776	16.31
Reference	Constant	0.5982	0.5000	17.05

sex from the frozen latent representations learned during Stage 1. The resulting performance metrics, summarised in Table 2, reveal clear and systematic differences across encoder architectures and training strategies.

For conventionally trained (Regular) models, both age and sex can be recovered from the ECG latent representation with performance substantially above chance, confirming that CLF information is implicitly encoded even when the encoder is trained solely for AFib classification. The extent of this CLF leakage depends strongly on model capacity. The high-capacity Attia encoder exhibits markedly stronger encoding of CLFs (Sex AUROC 0.8053; Age MAE 11.55) than the more compact ECGencode model (Sex AUROC 0.7196; Age MAE 14.48). This is consistent with larger latent spaces providing more representational freedom. Conversely, while compact architectures impose an implicit information bottleneck, they remain susceptible to CLF leakage and do not guarantee invariance.

Introducing GRe substantially reduces the recoverability of both age and sex across both encoder architectures. For both ECGencode and Attia, auxiliary prediction performance approaches near-random levels, indicating that the ECG latent representation no longer contains readily exploitable CLF information. The Balanced GRe configuration, which assigns equal loss weights across branches, yields the strongest suppression of CLF information. However, the standard GRe variant achieves a comparable reduction in leakage, suggesting that strict balancing is not strictly necessary to obtain effective CLF removal.

Overall, these results confirm that GRe successfully enforces CLF invariance in the ECG latent representation, independent of encoder capacity, and provide quantitative evidence that conventional ECG encoders implicitly encode CLFs unless explicitly constrained.

4.4. Synthetic multiplicative relationship under distribution shift

To isolate and illustrate the benefits of explicitly modelling multiplicative interactions, a synthetic proof-of-concept experiment is conducted. The objective of this experiment is not to replicate a physiological process, instead, it demonstrates a failure mode of the evaluated dense layers when learning multiplicative relations under distribution shift and contrasts this behaviour with the proposed MFID layer.

Let $\mathbf{z} = \{z_1, z_2, \dots\}$ denote latent features and $\mathbf{c} = \{c_1, c_2, \dots\}$ denote scalar conditioning variables. The models receive both vectors as input and are trained to predict one scalar target. The target variable Y is generated through an explicit multiplicative relationship of the form (3).

$$Y = \sum_{(i,j)\in S} z_i \cdot c_j \cdot w_{ij} \tag{3}$$

Table 3
Synthetic distribution-shift definitions used for the evaluation in Section 4.4. Latent features are denoted by z_k and conditioning variables (i.e., CLFs) by c_k.

Scenario	Training distribution	Testing distribution
Extreme	$z_k \sim U(0.0, 0.4)$ $c_k \sim U(0.0, 0.1)$	$z_k \sim U(0.6, 1.0)$ $c_k \sim U(0.3, 1.0)$
Realistic	$z_k \sim U(0.0, 1.0)$ $c_k \sim U(0.3, 0.7)$	$z_k \sim U(0.0, 1.0)$ $c_k \sim U(0.0, 1.0)$

Table 4
MSE scores for the synthetic multiplicative regression task, averaged over 10 independent runs. The *Extreme* scenario evaluates extrapolation under disjoint training and test ranges, while the *Realistic* scenario reflects partial coverage of the conditioning-variable domain during training. Lower test MSE indicates better generalisation of the underlying multiplicative relation under distribution shift.

Data set	Model	Train MSE	Test MSE
Extreme	Single Dense	$1.27 \times 10^{-5} \pm 6.36 \times 10^{-6}$	$2.34 \times 10^{-1} \pm 4.66 \times 10^{-3}$
	Three Dense	$\mathbf{2.63 \times 10^{-7} \pm 2.77 \times 10^{-7}}$	$1.81 \times 10^{-1} \pm 5.50 \times 10^{-2}$
	MFID LeakyReLU	$3.03 \times 10^{-7} \pm 6.29 \times 10^{-7}$	$\mathbf{1.96 \times 10^{-5} \pm 3.39 \times 10^{-5}}$
Realistic	Single Dense	$3.93 \times 10^{-3} \pm 9.63 \times 10^{-4}$	$2.08 \times 10^{-2} \pm 9.82 \times 10^{-4}$
	Three Dense	$\mathbf{1.82 \times 10^{-4} \pm 2.58 \times 10^{-4}}$	$3.93 \times 10^{-3} \pm 1.76 \times 10^{-3}$
	MFID LeakyReLU	$4.86 \times 10^{-4} \pm 1.01 \times 10^{-3}$	$\mathbf{4.91 \times 10^{-4} \pm 1.01 \times 10^{-3}}$

Here, S is the sparse set of active latent–conditioning pairs. Only a limited subset of pairs (z_i, c_j) is associated with a non-zero weight w_{ij}. This sparsity reflects the assumption that only certain features are relevant for multiplicative interactions. The reported MSE in Table 4 is computed between this ground-truth scalar Y and the model prediction $\hat{Y}$. No noise is added to Y, so the metric directly evaluates how well the model recovers and generalises the known multiplicative function.

To evaluate generalisation under distribution shift, two train/test range configurations are considered. The *Extreme* and *Realistic* scenarios are defined by the input ranges summarised in Table 3. The Extreme scenario uses 10,000 training samples and 2000 test samples, with two latent features and two conditioning variables. It contains two active pairs: z_1c_1 with weight 0.6 and z_2c_2 with weight 0.7. The Realistic scenario uses 100,000 training samples and 2000 test samples, with 128 latent features and two conditioning variables. Of the 256 possible pairwise interactions, only 12 are active, with weights ranging from 0.1 to 0.9; all remaining pairs have zero weight.

Table 4 compares single-layer and three-layer dense networks with Rectified Linear Unit (ReLU) activation against the proposed MFID layer in both scenarios. All experiments are repeated ten times with different random seeds to account for variability in training data sampling.

The results reveal a consistent failure mode for the evaluated dense architectures in these synthetic settings. While deeper dense networks achieve extremely low training error by tightly fitting the observed input domain, their test performance deteriorates substantially under distribution shift, particularly in the Extreme scenario, where the generalisation gap spans several orders of magnitude. In contrast, the MFID layer maintains low and stable error across both training and test distributions in both scenarios. This behaviour confirms that MFID learns the underlying multiplicative structure directly, rather than approximating it locally, and that its gating mechanism correctly suppresses spurious interactions when no multiplicative dependency is present.

4.5. AFib classification performance for downstream task

Table 5 summarises downstream AFib classification performance across the Detection and Prediction tasks. The evaluation is primarily intended to characterise the effect of controlled CLF information flow on performance and generalisation, rather than to optimise a single model configuration for peak performance in every task and cohort.

Accordingly, the results illustrate how invariance in the ECG latent representation, explicit CLF reintroduction, and task characteristics interact, clarifying when the proposed framework is most beneficial.

4.5.1. Impact of explicit CLF reintroduction

A consistent pattern emerges across encoders and tasks: the more effectively age and sex are removed from the ECG latent representation during Stage 1, the more critical their explicit reintroduction becomes during downstream classification. This also helps explain why explicitly reintroducing CLFs can substantially improve GRe-trained configurations in some settings: GRe suppresses latent CLF leakage, while MFID provides one mechanism for modelling how CLFs modulate ECG-derived evidence. When GRe is used without explicit CLF inputs, valuable CLF information is not available to the downstream classifier and reduced performance is expected. Conversely, when MFID is used without GRe, the explicit CLF pathway can be partially redundant because CLFs information may already be present in the ECG latent representation. This behaviour is most clearly illustrated by the high-capacity Attia encoder on the Prediction task.

The *Regular* Attia model, which implicitly encodes age and sex in its ECG latent representation, exhibits only marginal gains when CLFs are added explicitly through dense or MFID heads. In contrast, the *Balanced GRe* Attia model, where the ECG latent representation is invariant to these CLFs, shows a pronounced performance recovery when the CLFs are reintroduced. For instance, in the Prediction task, adding MFID increases the AUROC from 0.7003 to 0.7811 on CODE-15%, indicating that explicit reintroduction restores information that was no longer available through the ECG latent representation.

A similar, though less pronounced, effect is observed for the ECGencode architecture. Because ECGencode captures less CLF information even under regular training, the relative benefit of explicit CLF reintroduction is present but smaller, particularly for in-distribution evaluation.

4.5.2. Task-specific dynamics: Detection versus prediction

The relative importance of CLF integration differs substantially between Detection and Prediction tasks. For Detection of active AFib, a task that can be performed from the ECG alone, strong results are obtained even without explicit access to the CLFs.

In contrast, the Prediction task is more dependent on risk context, including the CLFs. In this setting, GRe-trained encoders without explicit CLFs perform worse, as expected, since the ECG latent representation has been deliberately stripped of age- and sex-linked information. Reintroducing CLFs through dense layers or MFID restores predictive performance, often to levels comparable with unconstrained baselines. For regularly trained models, the difference between variants with and without explicit CLFs is much smaller, further suggesting that these models already use age- and sex-linked information through the ECG latent representation.

This task-dependent behaviour aligns with clinical intuition and supports the view that the proposed framework controls how CLF information enters the model.

4.5.3. Generalisation under distribution shift

Evaluation on the external PTB-XL data set provides additional insight into robustness under distribution shift. Across configurations, changes are most consistently reflected in AUROC, which captures discrimination independent of a fixed operating threshold. By contrast, F1-scores depend on the chosen threshold and are therefore expected to be more sensitive to dataset shift.

While *Regular* models frequently achieve the highest performance on the in-distribution CODE-15% test set, they do not consistently generalise best to PTB-XL. Notably, for the Attia Prediction task, the *Balanced GRe + MFID* configuration attains a substantially higher OOD AUROC (0.7576) than the corresponding Regular baseline (0.6914). The corresponding *Balanced GRe + Dense* configuration reaches an AUROC of 0.7682.

F1-scores on PTB-XL are consistently lower than suggested by AUROC values. This discrepancy arises from the evaluation protocol: the classification threshold is optimised on the CODE-15% validation set and applied unchanged to PTB-XL. Given the distribution shift between data sets, this fixed threshold is likely suboptimal, and the resulting F1-scores should therefore be interpreted with caution.

Taken together, these results indicate that the proposed framework can improve robustness under population and acquisition shifts, even when in-distribution performance gains are limited. This behaviour is consistent with a trade-off between two competing factors: (i) the model's capacity to exploit cohort-specific correlations and CLF leakage present in the training distribution; and (ii) a controlled CLF pathway that supports interpretability and reduces reliance on such shortcuts. In this context, GRe constrains the encoder by suppressing CLF information in the ECG latent representation. This may reduce apparent performance on in-distribution data where CLFs are informative through uncontrolled latent shortcuts, while improving robustness when those correlations shift.

4.5.4. Age- and sex-stratified prediction performance

Because the largest model differences occur for the Prediction task, the Prediction results are further stratified by age band and sex. Fig. 3 summarises these results. The analysis compares Regular + MFID, B-GRe, and B-GRe + MFID: a conventional hybrid model with explicit age and sex, a model where age- and sex-linked information is suppressed without reintroduction, and the proposed controlled reintroduction pathway.

The age-stratified results support the main interpretation without showing a systematic loss in older patients. In the ≥ 75 subgroup, B-GRe + MFID remains close to Regular + MFID on CODE-15% and exceeds it on PTB-XL for both encoders. The youngest CODE-15% subgroup (< 50 years) shows a different pattern: B-GRe is comparable to, or slightly better than, the variants with explicit age and sex. This suggests that age is not always helpful in the same direction. For younger patients, explicit age information may sometimes lower the predicted risk, which can help explain why B-GRe remains competitive in this subgroup.

Sex-stratified results show the same broad trend in both male and female subgroups. In the all-age PTB-XL setting, B-GRe + MFID outperforms Regular + MFID in both sexes for both encoders. The fine-grained PTB-XL age-sex estimates are reported for transparency, but several cells contain few positive Prediction ECGs, especially below 50 years, and should be interpreted descriptively.

4.6. Model decision insights

The final component of the evaluation applies FIGI to analyse how training strategies affect how models integrate ECG signal characteristics and explicit CLFs in the decision-making process. By decomposing each prediction into attributions from the ECG latent representation and the explicitly provided CLFs, FIGI highlights that interpretation depends critically on how and where CLF information is encoded during training. The interpretation is therefore not based on the assumption that larger CLF attribution is inherently better, but on whether the model can still access CLFs through an uncontrolled latent pathway or only through the explicit CLF inputs.

Fig. 4 shows the distribution of modality-level importance scores, computed using IG, for the Prediction task on the test set. A pronounced contrast emerges between conventionally trained (*Regular*) models and balanced GRe + MFID models, particularly for the high-capacity Attia architecture as shown in Fig. 4.

In the *Regular* Attia model, the explicitly provided Age and Sex variables receive near-zero attribution, despite age being a well-established risk factor for AFib. In isolation, this pattern could be misread as evidence that CLFs play no role in the model's prediction. However, it

Table 5
Downstream AFib classification performance for Detection (active AFib) and Prediction (future AFib). Stage 1 encoder variants (Regular, GRe, and B-GRe for *Balanced* GRe) are evaluated with and without explicit CLF reintroduction in Stage 2 (Dense or MFID). In-distribution results are reported on CODE-15%, and out-of-distribution results are reported on PTB-XL. Bold values indicate the best performance within each encoder-task grouping.

Data variant	Encoder	Model variant	CODE-15% AUROC	CODE-15% F1	PTB-XL AUROC	PTB-XL F1
Detection	ECGencode	GRe	0.9750	0.6269	0.9774	**0.5968**
		GRe + Dense	0.9768	0.6326	0.9720	0.4809
		GRe + MFID	0.9712	0.6192	0.9561	0.4618
		B-GRe	0.9822	0.6273	**0.9786**	0.4564
		B-GRe + Dense	0.9835	0.6529	0.9728	0.3231
		B-GRe + MFID	0.9849	0.6511	0.9740	0.4257
		Regular	**0.9860**	**0.6780**	0.9742	0.3342
		Regular + Dense	0.9842	0.6720	0.9727	0.3930
		Regular + MFID	0.9830	0.6573	0.9690	0.5294
	Attia	GRe	0.9685	0.6866	0.9407	0.3593
		GRe + Dense	0.9791	0.7562	0.9584	0.4990
		GRe + MFID	0.9793	0.7511	0.9556	0.5098
		B-GRe	0.9212	0.4094	0.9618	0.5861
		B-GRe + Dense	0.9585	0.5039	0.9699	0.6877
		B-GRe + MFID	0.9596	0.5053	**0.9708**	0.6725
		Regular	**0.9810**	0.7880	0.9618	**0.6905**
		Regular + Dense	0.9807	**0.7910**	0.9354	0.4759
		Regular + MFID	0.9796	0.7825	0.9287	0.5144
Prediction	ECGencode	GRe	0.7997	0.1020	0.7235	0.0833
		GRe + Dense	0.8350	0.1080	0.7694	0.0675
		GRe + MFID	0.8272	0.1019	0.7556	0.0673
		B-GRe	0.7967	0.1005	0.7209	**0.0847**
		B-GRe + Dense	0.8360	0.1065	0.7700	0.0635
		B-GRe + MFID	0.8292	0.0948	**0.7803**	0.0460
		Regular	0.8142	**0.1191**	0.7340	0.0405
		Regular + Dense	0.8398	0.1099	0.7737	0.0471
		Regular + MFID	**0.8399**	0.1054	0.7280	0.0619
	Attia	GRe	0.7372	0.1229	0.6074	0.0000
		GRe + Dense	0.7671	0.1447	0.6507	0.0343
		GRe + MFID	0.7622	0.1520	0.6463	0.0382
		B-GRe	0.7003	0.0667	0.6965	0.0361
		B-GRe + Dense	0.7868	0.0863	**0.7682**	0.0274
		B-GRe + MFID	0.7811	0.0851	0.7576	0.0390
		Regular	0.7921	**0.1602**	0.6914	0.0331
		Regular + Dense	0.7925	0.1566	0.6339	0.0141
		Regular + MFID	**0.7934**	0.1581	0.6577	**0.0440**

is consistent with the CLF leakage quantified in Section 4.3: age and sex are already implicitly encoded within the ECG latent representation, rendering explicit inputs partially redundant. As a consequence, IG assigns little additional importance to these inputs, even when the underlying decision remains strongly influenced by CLFs.

In contrast, the *Balanced GRe + MFID* Attia model exhibits a fundamentally different attribution profile. With CLF information effectively removed from the ECG latent representation, the model is forced to rely on the explicitly provided CLFs to contextualise ECG-derived evidence. As a result, age contributes approximately 15–20% of the overall decision importance, with sex accounting for an additional 5–10% on average. This redistribution aligns with clinical intuition for prognostic tasks, where risk is determined by the interaction between ECG-derived evidence and CLFs, rather than by ECG signal characteristics alone.

Although these modality-level percentages are not directly comparable to coefficients in clinical risk scores, the relative ordering is qualitatively consistent with established scoring schemes. For example, the FHS score allocates a substantial fraction of the available points to age, whereas the attribution of sex is smaller (Schnabel et al., 2009). Similarly, $HARMS_2$-AF assigns a smaller but non-zero share of the total score to age and sex (Segan et al., 2023).

A similar, though less pronounced, trend is observed for the ECGencode architecture as shown in Fig. 4. Due to its more limited representational capacity, the Regular ECGencode model already exhibits reduced CLF leakage and therefore assigns non-zero importance to explicit age and sex inputs even without GRe training. Applying GRe further sharpens this separation, increasing the relative attribution of CLFs and clarifying their role in the final decision.

Interestingly, age contributes more to True Negative (TN) than to True Positive (TP) predictions in Fig. 4. This suggests that young age is often used as evidence against future AFib, rather than old age being used uniformly as evidence for future AFib. This is consistent with the age-stratified Prediction results, where models with explicit age and sex are not consistently superior in the youngest CODE-15% subgroup.

Finally, to ensure that enforcing CLF invariance does not compromise physiological interpretability, Grad-CAM++ heatmaps are generated on the raw ECG signal. As illustrated in Fig. 5, the model focuses on physiologically meaningful regions such as the absence of a P-wave, which is a diagnostic marker for active AFib. This indicates that GRe can suppress CLF shortcuts without eroding the morphological semantics required for arrhythmia detection.

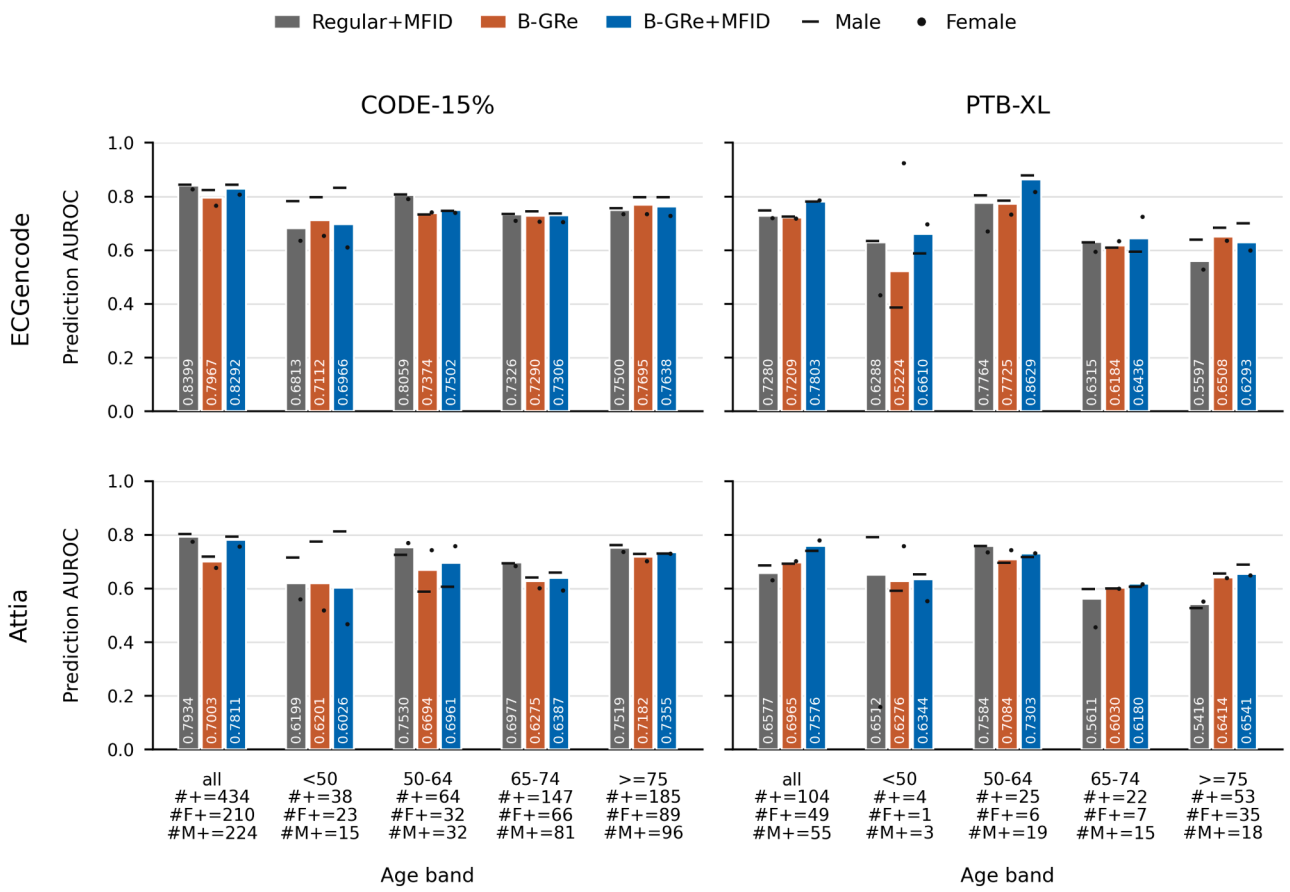


Fig. 3. Age- and sex-stratified AFib Prediction performance. Bars show sex-pooled AUROC by age band for Regular + MFID, B-GRe, and B-GRe + MFID. Horizontal ticks and points overlay male and female subgroup AUROCs, respectively, for the same model and age band. Values under each age band report the total, female, and male number of positive Prediction ECGs. PTB-XL fine-grained subgroup estimates, particularly below 50 years, are shown for completeness but should be interpreted descriptively because of limited positive counts.

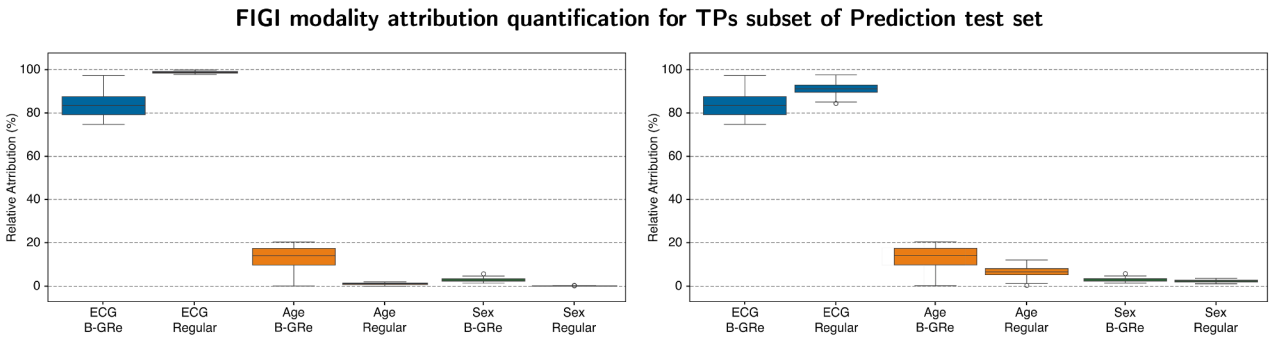
5. Discussion

Hybrid-AI approaches that combine CLFs (i.e., age and sex) with an ECG latent representation are motivated by the strong prognostic value of these CLFs in established clinical risk scores such as CHARGE-AF and $HARMS_2$-AF (Alonso et al., 2013; Segan et al., 2023). The central challenge highlighted by the present results is that, in DL-based ECG modelling, these CLFs are often implicitly encoded in the ECG latent representation. This CLF leakage is measurable using probe models and becomes more pronounced for higher-capacity encoders.

This implicit encoding has direct consequences for interpretation. When CLFs are appended via simple concatenation, attribution methods can assign the explicit CLF inputs little additional importance, even when the underlying CLF concepts are influential. This occurs because the same CLF information is already available through the ECG latent representation, making the explicit inputs partly redundant. In such cases, post-hoc analyses can incorrectly suggest an ECG-only decision process, obscuring the contribution of the CLFs to AFib prediction. The contrast between Regular and GRe-trained encoders, particularly for the higher-capacity Attia architecture, indicates that the apparent "unimportance" of explicit CLFs can be a symptom of representational redundancy rather than a clinical insight.

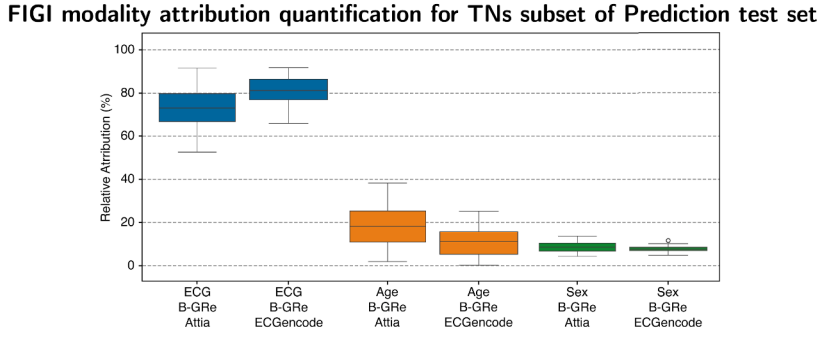
The proposed GRe–MFID–FIGI framework addresses this limitation by controlling *where* CLFs enter the model and how they interact with ECG latent features. GRe reduces CLF leakage, ensuring that age and sex influence the prediction predominantly through the explicit CLF pathway. This design makes the role of CLFs visible to IG-based modality attribution within FIGI, yielding importance patterns that better align with clinical expectations for prognostic tasks. In parallel, Grad-CAM++ heatmaps provide complementary evidence that physiologically meaningful ECG structures remain salient under CLF invariance. A limitation of this controlled two-stage design is that suppressing age and sex in the ECG latent representation may also suppress useful age- or sex-related ECG patterns. For example, PR-interval prolongation, reduced heart-rate variability, or an ECG-derived age may capture information that is not fully restored by reintroducing only biological age and sex. Thus, the second stage restores the intended explicit CLF pathway, but not necessarily every physiological signal removed during Stage 1. The subgroup analysis gives a first check on this limitation. It does not show a systematic degradation in older patients after controlled reintroduction. At the same time, the youngest CODE-15% subgroup shows that explicit access to age is not always beneficial, supporting the interpretation that age can sometimes act as evidence against future AFib. These subgroup patterns remain descriptive, especially for PTB-XL cells with few positive samples.

The MFID layer was introduced as a principled mechanism to model modulatory effects, in which CLFs influence how ECG latent features should be interpreted. The synthetic proof-of-concept confirms that explicitly parameterised multiplicative interactions generalise under distribution shift in settings where standard dense layers tend to approximate the relation only in-distribution. In the AFib experiments, the difference between MFID and conventional dense heads is smaller. This suggests that, for the evaluated cohorts and task formulations, either the relevant CLF–ECG interactions are largely captured by additive approximations or the data do not strongly probe regimes where explicit multiplicative structure is decisive.



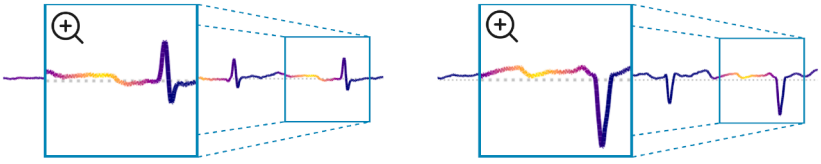
(a) Attia model variants (balanced and regular GRe)

(b) ECGencode model variants (balanced and regular GRe)



(c) Attia and ECGencode models (balanced GRe variant)

Fig. 4. FIGI modality attribution quantification. Boxplot analysis indicating the FIGI modality attribution quantification for the AFib Prediction task. Results are calculated using IG and averaged over CODE-15% test subsets across various model variants.



(a) Lead II (b) Lead V1

Fig. 5. Grad-CAM++ heatmaps on AFib ECG. Grad-CAM++ heatmaps, computed with respect to an intermediate convolutional layer, for a true-positive AFib Detection sample (Balanced GRe + MFID Attia model). Since the Attia encoder treats the 12 leads as input channels, the resulting map reflects a *temporal* region of importance rather than lead-specific importance.

MFID was introduced for this framework because it provides an explicit multiplicative path for reintroducing CLFs into the GRe latent space while retaining a conventional Dense path as an additive fallback. Its advantage in the synthetic proof-of-concept is relative to the evaluated Dense baselines but does not establish general superiority over other interaction mechanisms. Other interaction layers may also model these relations and could potentially outperform the configurations evaluated in this work.

Finally, the results indicate a trade-off between in-distribution performance and robustness. By reducing access to CLF shortcuts, GRe can constrain the encoder and reduce apparent in-distribution performance in settings where cohort-specific correlations are informative. However, improvements observed on the external PTB-XL evaluation suggest that enforcing CLF invariance in the ECG latent representation, followed by explicit CLF reintroduction, can promote representations that transfer better under population and acquisition shift. This robustness trend is intentionally reported conservatively and motivates further study, including evaluation of other classifiers within the same controlled framework.

6. Conclusion

This paper introduces Gradient-Reversed encoder (GRe), Multiplicative Feature-Interaction Dense (MFID), and Feature-Interaction Gradient Interpretation (FIGI). Together, these components form a unified hybrid-AI framework for ECG-based AFib modelling that explicitly controls CLF information flow. GRe enforces invariance to prespecified CLFs (age and sex in this study) in the ECG latent representation, mitigating CLF leakage and preventing CLF information from entering the model through uncontrolled latent shortcuts. MFID then reintroduces CLFs through gated multiplicative interactions with ECG latent features, providing an explicit pathway for modulatory effects. Finally, FIGI combines Grad-CAM++ localisation on the raw ECG with IG-based modality attribution on a latent proxy model to quantify, per input ECG, the relative attribution of the ECG latent representation versus CLFs.

The experiments demonstrate that CLF leakage can be substantial in conventionally trained encoders and tends to increase with model capacity. As a consequence, naive latent-space concatenation can obscure the role of CLFs: attribution quantification can suggest that the explicit

CLF inputs are unimportant because the same CLF information is already encoded implicitly in the latent ECG representation. By enforcing CLF invariance before reintroducing CLFs, more clinically plausible CLF attributions can be recovered without sacrificing signal-level interpretability. Age- and sex-stratified Prediction results further suggest that this controlled reintroduction does not cause a systematic performance loss in older patients, while showing that age can affect subgroups in different ways.

Several directions are important for future work. First, the robustness of FIGI could be improved through formal faithfulness testing, aggregation of modality attribution across multiple methods or bootstrap samples (e.g., comparing and averaging IG and DeepSHAP), and comparison of signal-localisation techniques (e.g., Grad-CAM++ versus SmoothGrad) to assess the stability of highlighted ECG regions. Second, the observed OOD performance trend for the proposed framework warrants further investigation across additional cohorts and shift settings. Such studies should also compare MFID with alternative feature-fusion or interaction mechanisms beyond the Dense heads evaluated here. Third, future extensions could reintroduce richer ECG-derived summaries alongside biological age and sex, such as ECG-derived age, to recover useful information that may be suppressed when enforcing age and sex invariance. Fourth, the different model variants could be combined for more informative risk assessment. If B-GRe, Regular+MFID, and B-GRe+MFID agree, the prediction is more likely to be supported by ECG morphology. If only the age- and sex-informed models agree, the prediction may depend more on clinical-factor context. The subgroup results suggest that this can work in both directions, since age may support some positive predictions while lowering predicted risk in younger patients. Defining clinically meaningful AFib subtypes from these patterns would require labels that are not available in the present data. Fifth, MFID could be augmented with regularisation strategies that discourage defaulting to purely additive behaviour and promote learning of genuine multiplicative relations. Finally, the computational efficiency of MFID could be improved by pruning interactions post-hoc, facilitating scaling to higher-dimensional latent spaces and broader clinical tasks.

Although the present study focuses on AFib, confounding by CLFs is pervasive in biomedical time series and medical imaging. The principles of invariant representation learning, explicit CLF reintegration, and modality-aware interpretation are therefore broadly applicable wherever patient characteristics shape both the measured signal and the clinical outcome.

Glossary:
AFib: Atrial Fibrillation
AI: Artificial Intelligence
AUROC: Area Under the Receiver Operating Characteristic Curve
CLF: Clinically Interpretable Feature
CVD: Cardiovascular Disease
DeepSHAP: Deep Shapley Additive Explanations
DL: Deep Learning
ECG: electrocardiogram
FIGI: Feature-Interaction Gradient Interpretation
GR: Gradient Reversal
GRe: Gradient-Reversed encoder
IG: Integrated Gradients
MAE: Mean Absolute Error
MFID: Multiplicative Feature-Interaction Dense
MSE: Mean Squared Error
NSR: Normal Sinus Rhythm
OOD: Out-of-Distribution
ReLU: Rectified Linear Unit
TN: True Negative
TP: True Positive

CRediT authorship contribution statement

Lennert Bontinck: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Karel Fonteyn:** Conceptualization, Writing – review & editing; **Antônio H. Ribeiro:** Conceptualization, Writing – review & editing; **Tom Dhaene:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing; **Dirk Deschrijver:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used OpenAI's ChatGPT in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Alonso, A., Krijthe, B. P., Aspelund, T., Stepas, K. A., Pencina, M. J., Moser, C. B., Sinner, M. F., Sotoodehnia, N., Fontes, J. D., Janssens, A. C. J. W., Kronmal, R. A., Magnani, J. W., Witteman, J. C., Chamberlain, A. M., Lubitz, S. A., Schnabel, R. B., Agarwal, S. K., McManus, D. D., Ellinor, P. T., Larson, M. G., Burke, G. L., Launer, L. J., Hofman, A., Levy, D., Gottdiener, J. S., Kääb, S., Couper, D., Harris, T. B., Soliman, E. Z., Stricker, B. H. C., Gudnason, V., Heckbert, S. R., & Benjamin, E. J. (2013). Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: The CHARGE–AF consortium. *Journal of the American Heart Association*, *2*(2). https://doi.org/10.1161/jaha.112.000102

Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., Erickson, B. J., Kapa, S., & Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *The Lancet*, *394*(10201), 861–867. https://doi.org/10.1016/S0140-6736(19)31721-0

Barros, A., German Mesner, I., Nguyen, N. R., & Moorman, J. R. (2024). Age prediction from 12-lead electrocardiograms using deep learning: A comparison of four models on a contemporary, freely available dataset. *Physiological Measurement*, *45*(8), 08NT01. https://doi.org/10.1088/1361-6579/ad6746

Bontinck, L., Dhaene, T., & Deschrijver, D. (2026). Chronological versus ECG-derived age in deep learning-based atrial fibrillation prediction. In *2026 IEEE EMBS International conference on biomedical and health informatics (BHI)* (p. 1–7). IEEE.

Bontinck, L., Fonteyn, K., Dhaene, T., & Deschrijver, D. (2024). Ecgencode: Compact and computationally efficient deep learning feature encoder for ecg signals. *Expert Systems with Applications*, *255*, 124775. https://doi.org/10.1016/j.eswa.2024.124775

Bontinck, L., Steyaert, A., Chen, H., Dhaene, T., & Deschrijver, D. (2025). Csd-afnet: Computationally efficient atrial fibrillation classification from ecgs using 2d causal strided dilated convolutions. In *2025 IEEE EMBS International conference on biomedical and health informatics (BHI)* (p. 1–7). IEEE. https://doi.org/10.1109/bhi67747.2025.11269513

Brant, L. C. C., Ribeiro, A. H., Eromosele, O. B., Pinto-Filho, M. M., Barreto, S. M., Duncan, B. B., Larson, M. G., Benjamin, E. J., Ribeiro, A. L. P., & Lin, H. (2025). Prediction of atrial fibrillation from the ECG in the community using deep learning: A multinational study. *Circulation: Arrhythmia and Electrophysiology*, *18*(10). https://doi.org/10.1161/circep.125.013734

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter conference on applications of computer vision (WACV)* (pp. 839–847). https://doi.org/10.1109/WACV.2018.00097

Chen, H., Lundberg, S. M., & Lee, S.-I. (2022). Explaining a series of models by propagating shapley values. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-31384-3

Cho, S., Eom, S., Kim, D., Kim, T.-H., Uhm, J.-S., Pak, H.-N., Lee, M.-H., Yang, P.-S., Lee, E., Attia, Z. I., Friedman, P. A., You, S. C., Yu, H. T., & Joung, B. (2024). Artificial intelligence-derived electrocardiographic aging and risk of atrial fibrillation: A multi-national study. *European Heart Journal*, *46*(9), 839–852. https://doi.org/10.1093/eurheartj/ehae790

Dong, T., Bao, X., Bi, J., & Chatterjee, S. (2025). Domain-adversarial pretrained encoder for ECG-based chagas disease screening. https://doi.org/10.22489/cinc.2025.091

Fonteyn, K., Bontinck, L., Dhaene, T., & Deschrijver, D. (2025). Augmentation-free longitudinal modeling through structuring whitened embeddings. *IEEE Access*, *13*, 159950–159960. https://doi.org/10.1109/ACCESS.2025.3608901

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd international conference on international conference on machine learning - volume 37* ICML'15 (pp. 1180–1189). JMLR.org.

Gliner, V., Makarov, V., Avetisyan, A. I., Schuster, A., & Yaniv, Y. (2023). Using domain adaptation for classification of healthy and disease conditions from mobile-captured images of standard 12-lead electrocardiograms. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-40693-6

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet. *Circulation*, *101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Hindricks, G., Potpara, T., Dagres, N., Arbelo, E., Bax, J. J., Blomström-Lundqvist, C., Boriani, G., Castella, M., Dan, G.-A., Dilaveris, P. E., Fauchier, L., Filippatos, G., Kalman, J. M., La Meir, M., Lane, D. A., Lebeau, J.-P., Lettino, M., Lip, G. Y. H., Pinto, F. J., Thomas, G. N., Valgimigli, M., Van Gelder, I. C., Van Putte, B. P., Watkins, C. L., & Group, E. S. D. (2020). 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the european association for cardio-thoracic surgery (EACTS): The task force for the diagnosis and management of atrial fibrillation of the european society of cardiology (ESC) developed with the special contribution of the european heart rhythm association (EHRA) of the ESC. *European Heart Journal*, *42*(5), 373–498. https://doi.org/10.1093/eurheartj/ehaa612

Jayakumar, S. M., Czarnecki, W. M., Menick, J., Schwarz, J., Rae, J. W., Osindero, S., Teh, Y. W., Harley, T., & Pascanu, R. (2020). Multiplicative interactions and where to find them. https://openreview.net/forum?id=rylnK6VtDH.

Lima, E. M., Ribeiro, A. H., Paixão, G. M. M., Ribeiro, M. H., Filho, M. M. P., Gomes, P. R., Oliveira, D. M., Sabino, E. C., Duncan, B. B., Giatti, L., Barreto, S. M., Meira, W., Schön, T. B., & Ribeiro, A. L. P. (2021a). Deep neural network estimated electrocardiographic-age as a mortality predictor. *medRxiv*. https://doi.org/10.1101/2021.02.19.21251232

Lima, E. M., Ribeiro, A. H., Paixão, G. M. M., Ribeiro, M. H., Pinto-Filho, M. M., Gomes, P. R., Oliveira, D. M., Sabino, E. C., Duncan, B. B., Giatti, L., Barreto, S. M., Meira, W., Jr, Schön, T. B., & Ribeiro, A. L. P. (2021b). Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-25351-7

Lin, C.-H., Liu, Z.-Y., Chu, P.-H., Chen, J.-S., Wu, H.-H., Wen, M.-S., Kuo, C.-F., & Chang, T.-Y. (2025). A multitask deep learning model utilizing electrocardiograms for major cardiovascular adverse events prediction. *npj Digital Medicine*, *8*(1). https://doi.org/10.1038/s41746-024-01410-3

Parraga, O., More, M. D., Oliveira, C. M., Gavenski, N. S., Kupssinskü, L. S., Medronha, A., Moura, L. V., Simões, G. S., & Barros, R. C. (2025). Fairness in deep learning: A survey on vision and language research. *ACM Computing Surveys*, *57*(6). https://doi.org/10.1145/3637549

Raghunath, S., Pfeifer, J. M., Ulloa-Cerna, A. E., Nemani, A., Carbonati, T., Jing, L., vanMaanen, D. P., Hartzel, D. N., Ruhl, J. A., Lagerman, B. F., Rocha, D. B., Stoudt, N. J., Schneider, G., Johnson, K. W., Zimmerman, N., Leader, J. B., Kirchner, H. L., Griessenauer, C. J., Hafez, A., Good, C. W., Fornwalt, B. K., & Haggerty, C. M. (2021). Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation–related stroke. *Circulation*, *143*(13), 1287–1298. https://doi.org/10.1161/CIRCULATIONAHA.120.047829

Rawshani, A., Rawshani, A., Smith, G., Boren, J., Bhatt, D. L., Börjesson, M., Engdahl, J., Kelly, P., Louca, A., Ramunddal, T., Andersson, E., Omerovic, E., Mandalenakis, Z., & Gupta, V. (2025). Integrating deep learning with ECG, heart rate variability and demographic data for improved detection of atrial fibrillation. *Open Heart*, *12*(1), e003185. https://doi.org/10.1136/openhrt-2025-003185

Ribeiro, A. H., Paixao, G. M. M., Lima, E. M., Horta Ribeiro, M., Pinto Filho, M. M., Gomes, P. R., Oliveira, D. M., Meira, W., Jr, Schon, T. B., & Ribeiro, A. L. P. (2021). CODE-15%: A large scale annotated dataset of 12-lead ECGs. https://doi.org/10.5281/zenodo.4916206

Ribeiro, A. H., & Ribeiro, A. L. P. (2024). Ai-ecg and prediction of new atrial fibrillation: When the heart tells the age. *European Heart Journal*, *46*(9), 853–855. https://doi.org/10.1093/eurheartj/ehae809

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira, W., Jr., Schön, T. B., & Ribeiro, A. L. P. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, *11*(1), 1760. https://doi.org/10.1038/s41467-020-15432-4

Schmitt, M. (2002). On the complexity of computing and learning with multiplicative neural networks. *Neural Computation*, *14*(2), 241–301. https://doi.org/10.1162/08997660252741121

Schnabel, R. B., Sullivan, L. M., Levy, D., Pencina, M. J., Massaro, J. M., D'Agostino, R. B., Newton-Cheh, C., Yamamoto, J. F., Magnani, J. W., Tadros, T. M., Kannel, W. B., Wang, T. J., Ellinor, P. T., Wolf, P. A., Vasan, R. S., & Benjamin, E. J. (2009). Development of a risk score for atrial fibrillation (framingham heart study): A community-based cohort study. *The Lancet*, *373*(9665), 739–745. https://doi.org/10.1016/s0140-6736(09)60443-8

Segan, L., Canovas, R., Nanayakkara, S., Chieng, D., Prabhu, S., Voskoboinik, A., Sugumar, H., Ling, L.-H., Lee, G., Morton, J., LaGerche, A., Kaye, D. M., Sanders, P., Kalman, J. M., & Kistler, P. M. (2023). New-onset atrial fibrillation prediction: The HARMS2-AF risk score. *European Heart Journal*, *44*(36), 3443–3452. https://doi.org/10.1093/eurheartj/ehad375

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. https://doi.org/10.48550/ARXIV.1312.6034

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., & Wattenberg, M. (2017). Smoothgrad: Removing noise by adding noise. *CoRR*, *abs/1706.03825*. http://arxiv.org/abs/1706.03825.

Strodthoff, N., Lopez Alcaraz, J. M., & Haverkamp, W. (2024). Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: An explorative study in emergency care. *European Heart Journal - Digital Health*, *5*(4), 454–460. https://doi.org/10.1093/ehjdh/ztae039

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th international conference on machine learning - volume 70* ICML'17 (pp. 3319–3328). JMLR.org.

Tsiartas, E., Nayak, D., & Meade, A. (2025). Artificial intelligence capabilities in identifying atrial fibrillation using baseline sinus rhythm ECG: A systematic review. *Open Heart*, *12*(2), e003657. https://doi.org/10.1136/openhrt-2025-003657

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, *7*(1), 154. https://doi.org/10.1038/s41597-020-0495-6

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Samek, W., & Schaeffter, T. (2022). Ptb-xl, a large publicly available electrocardiography dataset. https://physionet.org/content/ptb-xl/1.0.3/. https://doi.org/10.13026/KFZX-AW45

Wu, Y., Zhang, S., Zhang, Y., Bengio, Y., & Salakhutdinov, R. (2016). On multiplicative integration with recurrent neural networks. In *Proceedings of the 30th international conference on neural information processing systems* NIPS'16 (p. 2864–2872). Red Hook, NY, USA: Curran Associates Inc.

Yang, Y., Lin, M., Zhao, H., Peng, Y., Huang, F., & Lu, Z. (2024). A survey of recent methods for addressing AI fairness and bias in biomedicine. *Journal of Biomedical Informatics*, *154*, 104646. https://doi.org/10.1016/j.jbi.2024.104646

Yuan, N., Duffy, G., Dhruva, S. S., Oesterle, A., Pellegrini, C. N., Theurer, J., Vali, M., Heidenreich, P. A., Keyhani, S., & Ouyang, D. (2023). Deep learning of electrocardiograms in sinus rhythm from US veterans to predict atrial fibrillation. *JAMA Cardiology*, *8*(12), 1131. https://doi.org/10.1001/jamacardio.2023.3701