

1 An Interpretable Block-Term Tensor Network for 2 Incident Atrial-Fibrillation Prediction from the 3 Sinus-Rhythm 12-Lead ECG and Clinical Risk Factors

4 Axel Faes^{a,*}, Liesbet Peeters^a

^a*Data Science Institute and Biomedical Research Institute (BIOMED), University of Hasselt, Belgium*

5 Abstract

Deep-learning models predict incident atrial fibrillation (AF) from the sinus-rhythm 12-lead electrocardiogram (ECG) accurately but opaquely, which limits clinical use. We present the Block-Term Tensor Network (BTTN), a survival model that represents each ECG as a frequency-by-time-by-lead spectrogram tensor and factorizes it into blocks with interpretable per-block frequency and lead factors, trained under a Cox partial-likelihood loss. Age, sex, and ECG-interval risk factors enter through a zero-initialized linear term that bypasses the block layer, leaving the interpretable factors and the faithfulness probe unchanged. We use gold ICD-10 labels from MIMIC-IV, patient-level cross-validation, and external validation on CODE-15. The ECG-only model matched a black-box CNN (one-year AUROC 0.747 versus 0.734) but trailed a clinical baseline of age, sex, and ECG intervals (0.760). Fusing the ECG substrate with these risk factors reached 0.778 [0.767, 0.789] at one year, exceeding both the clinical baseline (paired +0.019 [+0.014, +0.024]) and the CNN, and 0.794 at five years. A linear probe on the interpretable factors recovered 97.6% of full-model discrimination, and an intervention test confirmed the model relies on its top-ranked factors. An interpretable ECG-plus-clinical model matched black-box discrimination here while giving measured, factor-level evidence for each prediction.

6 *Keywords:* Atrial fibrillation, electrocardiography, interpretable machine
7 learning, block-term tensor decomposition, survival analysis, deep learning

*Corresponding author.

Email addresses: `axel.faes@uhasselt.be` (Axel Faes),
`liesbet.peeters@uhasselt.be` (Liesbet Peeters)

8 1. Introduction

9 Atrial fibrillation is the most common sustained cardiac arrhythmia and a
10 leading cause of stroke, heart failure, and hospitalization. Much of its burden
11 is silent. Paroxysmal episodes come and go, and many patients first present
12 with a stroke before any documented arrhythmia. Oral anticoagulation re-
13 duces stroke incidence once AF is detected, so the bottleneck is detection
14 rather than treatment. Intermittent clinical ECGs miss paroxysmal episodes,
15 and prolonged ambulatory or implanted monitoring is too costly to deploy
16 broadly. A model that flags high-risk patients from a single resting 12-lead
17 ECG could focus monitoring on those most likely to benefit. The resting
18 ECG is cheap, ubiquitous, and already part of routine care.

19 Deep neural networks predict AF risk from the sinus-rhythm ECG. Attia
20 et al. [1] showed that a convolutional network identifies AF from an ECG
21 recorded during normal sinus rhythm within a narrow time window, and
22 later work extended the approach to future onset. Raghunath et al. [2]
23 predicted new-onset AF within one year at an AUROC near 0.85 with age
24 and sex added, and the Montreal Heart Institute ECG-AI model reached an
25 externally validated 0.77 for five-year incident AF on MIMIC-IV [3]. The
26 sinus-rhythm ECG carries a recoverable signal of latent atrial disease.

27 These accurate models are black boxes. A clinician cannot inspect the
28 risk score, audit it for shortcut learning, or reconcile it with physiological
29 understanding. Post-hoc explanation methods such as saliency maps and
30 Grad-CAM are widely used, but they approximate a trained network from
31 the outside. The approximation can diverge from the computation the net-
32 work performs, and its faithfulness is rarely measured. An explanation can
33 therefore misstate the reasoning behind a screening decision that triggers
34 long-term monitoring or therapy.

35 What the literature lacks is not another explanation method but an ac-
36 curate incident-AF model whose own interpretable representation is shown,
37 by direct measurement, to carry the decision: the accurate models are con-
38 volutional black boxes explained after the fact, and the faithfulness of those
39 explanations is rarely tested. We fill that gap by building the model around
40 interpretable block-term frequency and lead factors and then measuring how
41 much of its accuracy those factors account for. This is what motivates BTTN,
42 whose block-term structure keeps the per-block signatures readable while a
43 learnable nonlinear survival head supplies the capacity a linear model lacks.

44 We ask whether a model can be accurate and faithfully interpretable for

incident-AF prediction. The starting point is block-term tensor regression (BTTR) [4], a supervised tensor method whose per-block factors form an interpretable, low-rank description of the input. BTTR is linear, and a linear model cannot match a convolutional network on this task. We generalize the block-term structure into a learnable, nonlinear, survival-trained architecture, the Block-Term Tensor Network (BTTN). It keeps the interpretable per-block frequency and lead factors and adds the capacity to compete with a CNN. The ECG-only model matches the black-box CNN and falls short of a clinical baseline built from age, sex, and ECG intervals. The ECG and the risk factors carry partly complementary information. Age, sex, and the same ECG-interval measurements enter through a zero-initialized linear term added to the head log-risk that bypasses the block layer. The per-block signatures and the faithfulness probe stay unchanged, and the demographics cost nothing in interpretability. The fused model significantly beats both the clinical baseline and the black-box CNN.

The contributions are as follows. Incident-AF prediction is cast as a time-to-event problem on an alignment-free spectrogram tensor, and BTTN is trained end to end under a Cox partial-likelihood loss with the interpretable block-term factors intact. On gold ICD-10 labels, under patient-level cross-validation with patient-bootstrap confidence intervals, the interpretable ECG substrate fused with clinical risk factors through a transparent additive term significantly outperforms both the risk-factor baseline and a matched black-box CNN, and a single survival model gives stable risk from 90 days to five years. Faithfulness is measured directly, through the fraction of full accuracy a linear readout of the interpretable factors recovers and an intervention test on the trained model. The learned frequency, lead, and cardiac-phase substrate is clinically coherent and reproduces on an independent external cohort.

2. Related Work

2.1. ECG deep learning for incident AF

Deep learning on the sinus-rhythm ECG splits into two tasks that are easily conflated. One is identification of AF concurrent or near-concurrent with the recording: Attia et al. [1] flagged patients in or near an episode at AUROC about 0.87 within a one-month window, but that label window lets a model exploit signatures of an arrhythmia already present. The task we address is genuine future onset in a patient who is currently arrhythmia-free.

Several groups predict true onset. Raghunath et al. [2] reached one-year AUROC near 0.85 with demographic covariates and a lower figure on the ECG alone; Christopoulos et al. [5] reported an AI-ECG probability that rose with time to incident AF; and Gadaleta et al. [6] extended prediction to ambulatory single-lead recordings, so the signal is not confined to the resting 12-lead. Two studies treat the ECG and clinical risk as complementary: Khurshid et al. [7] found an ECG score and the CHARGE-AF risk factors [8] largely additive for five-year AF, and Jabbour et al. [3], the closest comparator on our data source, reported a five-year model with internal AUROC near 0.78 and external 0.77 on MIMIC-IV. The frontier for true incident AF thus sits at roughly 0.77 to 0.85, dominated by convolutional black boxes whose explanations, when offered, are post-hoc; our aim is an interpretable architecture whose factor-level reasoning is measured, cast as a survival model. AF detection from an ECG already in fibrillation is a separate, largely solved problem (AUROC near 0.96 to 1.0) and is not a comparator here.

Two recurring choices shape our design. Most incident-AF networks are fixed-horizon binary classifiers, which discard event timing and need a separate model per horizon; a time-to-event formulation uses the censoring structure directly and yields one risk function readable at any horizon, standard in clinical survival modeling but still uncommon in ECG deep learning. The deep-survival machinery is established: DeepSurv [9] trains under the Cox partial likelihood, Gensheimer and Narasimhan [10] bin the time axis for a likelihood-based hazard, and DeepHit [11] models competing risks, the lineage of our gold-ready variant (Appendix Appendix A). We take the Cox route on the spectral tensor and keep the interpretable block factors intact through it. The second choice is the input representation: median-beat and single-cycle representations align to the R-peak and average across beats, suppressing the rhythm and fibrillatory-band content that distinguishes atrial substrate, so we preserve spectral information across the full strip.

2.2. Interpretability in ECG-AI

Interpretability for ECG models is dominated by post-hoc attribution. Saliency, Grad-CAM, and SHAP highlight input regions that influence a trained network and have been used to argue that ECG classifiers attend to plausible waveform segments. They share a structural limit: they explain a fixed black box from the outside, and the explanation can diverge from the computation the model performs. Adebayo et al. [12] showed that popular saliency maps are insensitive to randomizing the weights or the labels, and

118 Jacovi and Goldberg [13] argue that faithfulness, the degree to which an ex-
119 planation reflects the model’s true reasoning, must be defined and measured
120 on a graded scale.

121 Our linear-readout test inherits the discipline of the probing literature.
122 Hewitt and Liang [14] show that a probe’s accuracy is interpretable only
123 against a control task, since a sufficiently expressive probe can recover struc-
124 ture the representation does not itself encode. Accordingly, the recovery fig-
125 ure is read as the fraction of full-model accuracy a deliberately *linear* readout
126 reaches, paired with the control probes in Section 4.

127 By-design interpretability is the closer relative. Concept-bottleneck archi-
128 tectures [15] build interpretability into the model, but information can leak
129 through the predictor head, leaving the interpretable layer as not the true
130 locus of the decision [16]. The most closely related work, the compositional
131 multilinear model of Doods et al. [17], like BTTN builds interpretability
132 from low-rank multilinear structure. BTTN departs from it in two ways: the
133 structure is cast as a Cox survival model for dated outcomes, and faithfulness
134 is a measured quantity, reporting how much of the trained model’s accuracy a
135 linear probe on its own factors recovers. The contribution here is that direct,
136 quantitative test of how much of the model’s accuracy its own interpretable
137 factors carry.

138 2.3. *Tensor methods and the block-term lineage*

139 Tensor decompositions factor a multiway array into a few low-rank com-
140 ponents, and the canonical polyadic and Tucker forms together with their
141 applications are surveyed by Kolda and Bader [18]. In signal processing and
142 machine learning these tools recover interpretable multiway structure and
143 compress high-dimensional models, a line traced by Cichocki et al. [19].
144 Block-term tensor decomposition [20] represents a tensor as a sum of low-
145 multilinear-rank terms, interpolating between the canonical polyadic and
146 Tucker models; applied to biomedical signals, it has separated physiologically
147 meaningful spatio-temporal sources, for example the block-term modelling of
148 epileptic EEG by Hunyadi et al. [21], which motivates our use of per-block
149 factors as readable frequency-and-lead signatures that can be inspected di-
150 rectly. Supervised block-term tensor regression [4] uses this structure for
151 prediction, with each block’s factor matrices forming an interpretable, low-
152 rank signature of the input modes. The strength of BTTR for our purpose
153 is this interpretable structure. Its weakness is that it is linear and capacity-
154 limited. A separate line of work uses block-term structure to compress the

weight tensors of neural layers; Ye et al. [22] factor recurrent-layer weights with a block-term decomposition to cut parameter counts, but the resulting factors are a means to compression and carry no interpretive intent. BTTN shares the block-term name with this compression work but not the goal: that line factors the model’s weights and reads nothing from the factors, whereas we factor the input and keep A_k , C_k as named physiological signatures while making the surrounding computation nonlinear and survival-trained. Self-supervised ECG representation learning, including contrastive and masked-prediction approaches, provides a way to use large unlabeled ECG pools. An in-house contrastive scheme runs on the block encoder, so the pretrained weights transfer exactly into the supervised model.

3. Methods

BTTN turns a resting ECG into a dated risk in six stages, shown end to end in Fig. 1. A sinus-rhythm 12-lead ECG (1) is converted lead by lead into a frequency-by-time-by-lead spectrogram tensor (2); each block projects that tensor onto an interpretable per-block frequency factor and lead factor (3); a multi-scale temporal convolution bank with attention pooling extracts the temporal pattern of each block’s response (4); age, sex, and ECG-interval risk factors enter through a transparent additive term that bypasses the block layer (5); and a Cox partial-likelihood head reads a relative hazard at any horizon (6). The subsections below detail each stage. The interpretable per-block frequency and lead factors are preserved throughout, and because the clinical term is added after the block layer, fusion leaves those factors and the faithfulness probe unchanged.

3.1. Cohort and incident-AF labels

The cohort is longitudinal, built from MIMIC-IV-ECG [23], which contains roughly 800,000 12-lead recordings from about 161,000 patients. The prediction unit is one ECG per patient. For each patient, the earliest sinus-rhythm, AF-naïve recording serves as the index ECG. A 30-day blanking period after the index ECG is excluded: it removes onsets concurrent with or immediately adjacent to the index recording, where a positive label would reflect a paroxysmal episode already in progress, keeping the task one of prediction rather than near-concurrent identification. A patient is positive if the first AF occurs in the interval from index plus 30 days to index plus the horizon, and is a control if no AF is ever observed and the patient is

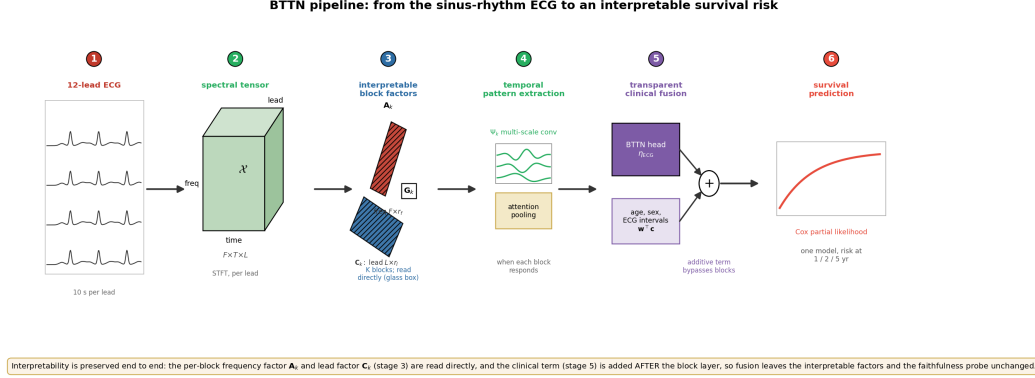


Figure 1: End-to-end BTTN pipeline. (1) A sinus-rhythm 12-lead ECG is (2) converted lead by lead into a frequency-by-time-by-lead spectrogram tensor, (3) projected by each block onto an interpretable per-block frequency factor A_k and lead factor C_k , (4) passed through a multi-scale temporal convolution bank and attention pooling, (5) fused with age, sex, and ECG-interval risk factors through a transparent additive term that bypasses the block layer, and (6) read as a Cox survival relative hazard at any horizon. Interpretability is preserved end to end: the per-block factors are read directly, and because the clinical term is added after the block layer, fusion leaves the interpretable factors and the faithfulness probe unchanged.

190 followed AF-free for at least the horizon. The primary horizon is 365 days,
 191 with secondary horizons at 730 and 1825 days. The outcomes are dated and
 192 right-censored, so the problem is naturally time-to-event.

193 Outcomes are ascertained from gold diagnosis codes in the MIMIC-IV
 194 hospital module. A patient is assigned a first-AF date from the earliest hos-
 195 pital record carrying an ICD-10 I48 code (atrial fibrillation and flutter) or
 196 an ICD-9 4273x code, linked to MIMIC-IV-ECG through the shared patient
 197 identifier. The dated diagnosis is the outcome from which both the event indi-
 198 cator and the time-to-event are computed. Sinus rhythm at index is read from
 199 the machine-measurement cart-report fields released with MIMIC-IV-ECG:
 200 a recording is treated as sinus when its automated interpretation matches
 201 **sinus** and is not flagged as AF, and a non-sinus filter (paced, junctional,
 202 atrial tachycardia, supraventricular tachycardia, wandering pacemaker, or
 203 ectopic atrial rhythm) disqualifies recordings whose rhythm is neither clean
 204 sinus nor AF, so that an ambiguous-rhythm ECG never serves as an AF-naive
 205 sinus index.

206 The index ECG is the earliest recording that is both sinus and not non-
 207 sinus per these filters, and “AF-naive at index” is enforced two ways: no

gold I48 diagnosis date may fall at or before index plus the 30-day blanking window, and no AF-flagged recording may exist at or before that same point. Restricting each patient to a single index ECG, the earliest eligible one, prevents the same patient from contributing correlated recordings to both training and test folds and removes the optimism that repeated within-patient ECGs would otherwise introduce.

Two cohort definitions are used. The discrimination experiments use a binary cohort at the primary one-year horizon with a fixed negative-to-positive ratio, which gives a controlled comparison of models at a single horizon. The survival experiments use the full right-censored cohort with continuous time-to-AF, which supports the multi-horizon and concordance analyses and the clinical-utility read-outs. In both cases the index ECG, the 30-day blanking period, and the patient-level split are identical, so the two cohorts differ only in how the outcome is encoded and not in which patients are eligible.

The eligibility flow is as follows. Of the patients in MIMIC-IV-ECG who carry a gold I48 diagnosis date in the hospital module, 35,167 have such a date; we exclude 14,469 patients with prevalent or concurrent AF, whose first I48 code falls at or before index plus the blanking window. At the primary one-year horizon this leaves a cohort of 17,855 patients with 1,855 incident-AF cases (Table 1). Against the full pool of controls followed AF-free through one year, the natural one-year incidence is $1,855 / (1,855 + 41,770) = 4.25\%$, which is the base rate used for the true-incidence recalibration in the clinical-utility read-out. For discrimination, the cohort is class-balanced by down-sampling controls to a four-to-one negative-to-positive ratio, giving a modelling cohort of $n = 9,275$ (1,855 cases and 7,420 controls). This balancing raises the apparent prevalence to 0.20, so the operating-point statistics on this cohort are reported after recalibration to the 4.25% natural incidence (Section 4).

3.2. Spectral tensorization

A common ECG representation for tensor methods is the R-peak-aligned median beat, but median-beat averaging discards rhythm and fibrillatory-band information, and a linear block-term model cannot beat a CNN on it. The input here is an alignment-free spectral tensor. Each 10-second strip is denoised per lead with a high-pass baseline-drift filter (second-order Butterworth, 0.5 Hz cutoff), a 50 Hz powerline notch ($Q = 30$), and a `sym4` discrete-wavelet soft threshold at level 3 (threshold $7.0 \times$ the detail-coefficient

Table 1: Cohort accounting for the gold ICD-10 MIMIC-IV analysis (one index ECG per patient). The modelling cohort is the class-balanced subset used for discrimination; operating points are recalibrated to the natural 4.25% incidence.

Gold ICD-10 accounting	Patients
Patients with a gold I48 diagnosis date	35,167
less prevalent/concurrent AF	−14,469
One-year cohort	
Patients	17,855
Incident-AF cases	1,855
Natural 1-yr incidence	4.25%
Modelling (balanced) cohort	
Cases	1,855
Controls (downsampled, 4:1)	7,420
Total n	9,275
Apparent prevalence	0.20

standard deviation), then max-abs normalized and decimated by a factor of two from the native 500 Hz to 250 Hz, which preserves the sub-40 Hz P-wave and atrial band. For each lead, the magnitude short-time Fourier transform is then computed (`scipy.signal.spectrogram`, segment length `nperseg`=128, overlap 96 so the hop is 32 samples, at the decimated 250 Hz rate), the frequency bins up to 40 Hz are kept, a $\log(1 + x)$ magnitude compression is applied, the time mode is linearly resampled to a fixed $T = 40$ frames, and each sample is standardized by its own tensor-wide mean and standard deviation. The fold-level standardization that precedes the network (subtracting a mean and dividing by a standard deviation computed over the frequency, time, and lead axes) is estimated on the training fold only and applied unchanged to the held-out fold, so no test statistics leak into training. Stacking the 12 leads yields

$$\mathcal{X} \in \mathbb{R}^{F \times T \times L}, \quad F \approx 21, \quad T = 40, \quad L = 12, \quad (1)$$

with F spanning 0 to 40 Hz, T the resampled time mode, and L the leads. A magnitude spectrogram is tolerant to small temporal shifts, which removes the beat-phase misalignment that makes block-term regression fail on the raw time strip, without averaging individual beats away.

262 3.3. The Block-Term Tensor Network

263 BTTN keeps the interpretable block structure of block-term regression
 264 and makes it a learnable nonlinear layer trained by backpropagation. The
 265 layer comprises K blocks. Block k holds a frequency factor $A_k \in \mathbb{R}^{F \times r_f}$, a
 266 per-block spectral signature; a lead factor $C_k \in \mathbb{R}^{L \times r_l}$, a per-block spatial
 267 signature; and a multi-scale temporal convolution bank Ψ_k , implemented
 268 as grouped one-dimensional convolutions with kernel sizes 3, 7, and 15. The
 269 convolution bank is the learnable, local replacement for the static score vector
 270 of linear block-term regression.

271 The forward pass proceeds in three stages. First, each block projects the
 272 frequency and lead modes of the input,

$$\mathcal{Z}_k = \mathcal{X} \times_1 A_k^\top \times_3 C_k^\top, \quad (2)$$

273 where $\times_n$ denotes the mode- n product, producing a reduced tensor of shape
 274 (r_f, T, r_l) per block. Second, the block tensors are reshaped to $K r_f r_l$ channels
 275 over the time mode, passed through the multi-scale grouped convolutions,
 276 concatenated across scales, and followed by ReLU and batch normalization.
 277 Third, the layer applies temporal attention pooling,

$$\mathbf{a} = \text{softmax}_T(\text{conv}_{1 \times 1}(\mathcal{Z})), \quad \mathbf{z}_{\text{att}} = \sum_t \mathbf{a}_t \odot \mathcal{Z}_t, \quad (3)$$

278 where $\odot$ is elementwise multiplication, giving an interpretable account of
 279 when in the recording each block responds, and this is concatenated with a
 280 global max-pool. A multilayer perceptron head with batch normalization,
 281 dropout, and ReLU maps the pooled representation to a scalar ECG log
 282 relative hazard.

283 Clinical risk factors enter through a transparent additive term rather
 284 than through the block layer. The fused log relative hazard is $\eta = \eta_{\text{ECG}} +$
 285 $\mathbf{w}^\top \mathbf{c}$, where η_{ECG} is the BTTN head output and $\mathbf{w}^\top \mathbf{c}$ is a linear function
 286 of the covariate vector $\mathbf{c}$. The covariate term uses the same age, sex, and
 287 ECG-interval features as the clinical baseline; its weights $\mathbf{w}$ are warm-started
 288 from a logistic fit on those features, initialized so that the term contributes
 289 zero at the start of training, and exempted from weight decay. The term
 290 is added to the head log-risk and never touches the block projection. The
 291 per-block frequency factor A_k and lead factor C_k , the attention weights, and
 292 the faithfulness probe are all unchanged by fusion. The demographics are

293 added at no interpretability cost, and the interpretable-substrate read-outs
 294 below are computed on the ECG term alone.

295 The block factors A_k and C_k are the interpretable maps. They define
 296 the frequency-by-lead subspace each block can access, and their outer prod-
 297 uct $A_k \otimes C_k$ gives a per-block frequency-by-lead signature that can be read
 298 directly. Defining the accessible subspace does not guarantee that the sub-
 299 space carries the decision. Whether it does is an empirical question, answered
 300 here by measuring probe recovery. Figure 2 contrasts the linear block-term
 301 regressor with BTTN.

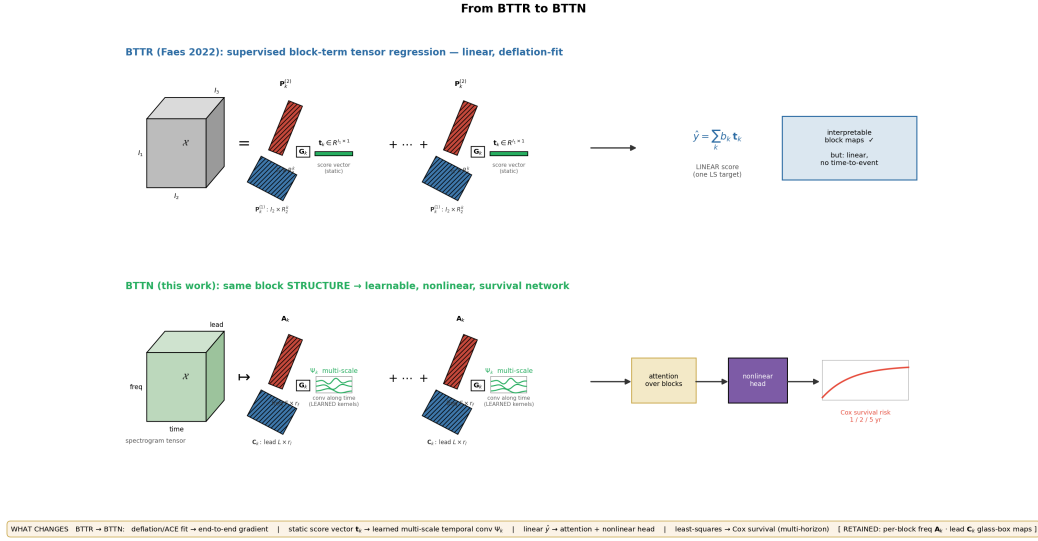


Figure 2: Method schematic. Linear block-term tensor regression projects the input onto per-block factors and combines them with a static score, whereas BTTN keeps the interpretable per-block frequency factor A_k and lead factor C_k but replaces the static score with a multi-scale temporal convolution bank and attention pooling, trained end to end with a Cox survival loss.

302 3.4. Survival objective

303 The network output is a log relative hazard, and we train it with the Cox
 304 partial negative log-likelihood [24] using within-batch risk sets,

$$\mathcal{L} = -\frac{1}{\sum_i \delta_i} \sum_i \delta_i \left(\eta_i - \log \sum_{j: t_j \geq t_i} e^{\eta_j} \right), \quad (4)$$

305 where η_i is the predicted risk for patient i , δ_i is the event indicator, and t_i
 306 is the time to event or censoring. The risk set $\{j : t_j \geq t_i\}$ is approximated
 307 within each mini-batch: the batch is sorted by descending time and the de-
 308 nominator evaluated with a log-cumulative-sum-of-exponentials, so the par-
 309 tial likelihood is computed against the batch-local risk set. This approxima-
 310 tion is the reason the batch size is held large at 512; a small batch would give
 311 a high-variance risk-set estimate and bias the gradient, and any batch con-
 312 taining no events is skipped so the normalizer is well defined. Optimization
 313 uses Adam (learning rate 2×10^{-3} , weight decay 10^{-4}) with a cosine-annealing
 314 schedule over the maximum of 120 epochs. Early stopping monitors AUROC
 315 at the prediction horizon on an inner validation split, an 85/15 patient-level
 316 partition of the *training* fold drawn before any test data is touched, with a
 317 patience of 16 epochs; the test fold is never used for model selection. Three
 318 independently seeded models are ensembled per fold to match the CNN base-
 319 line. A single trained model yields risk at any horizon, so one survival model
 320 produces the one-, two-, and five-year readouts.

321 For settings with gold date-of-death information, a discrete-time competing-
 322 risks variant is implemented in the style of a cause-specific deep survival
 323 model [11]. It addresses the bias that arises when elderly patients die before
 324 AF, which naive right-censoring would otherwise treat as if AF were still
 325 possible. The variant is an implementation sanity check: it runs correctly on
 326 synthetic competing-risks data (AF-cause C-index 0.867) and is gold-ready,
 327 awaiting real date-of-death information. It has not been run on the present
 328 cohort, so the censoring bias it targets is not mitigated in the numbers re-
 329 ported here, and its architecture is given in Appendix Appendix A.

330 3.5. Self-supervised pretraining

331 The labeled set is small and label-noisy, while the unlabeled ECG pool
 332 is large. The same block-term layer is pretrained with a SimCLR-style con-
 333 trastive objective on roughly 30,000 unlabeled spectral tensors, using time
 334 masking, frequency masking, lead dropout, additive noise, and small time
 335 rolls as augmentations and an NT-Xent loss. Because the contrastive en-
 336 coder is the block layer itself, the pretrained weights initialize the supervised
 337 BTTN exactly, transferred into the block factors A_k , C_k and the temporal-
 338 convolution bank by a partial state-dict load. Pretraining is label-free and
 339 acts on the ECG substrate only, independent of the gold-label supervision.

340 The architecture and training settings are collected in Table 2, with all
 341 values taken from the released configuration and model code.

Table 2: BTTN architecture and training hyperparameters (from the released code). The block count K is swept separately (Section 4); the value below is the default.

Setting	Value
Blocks K	12 (default; swept 4–24)
Frequency factor rank r_f	3
Lead factor rank r_l	2
Temporal conv kernels	3, 7, 15 (grouped, multi-scale)
Conv channel multiplier	6 per group
Temporal pooling	attention (1×1 conv $\rightarrow$ softmax) + max
MLP head width	256 ($\rightarrow$ 128 $\rightarrow$ 1)
Dropout	0.4
Optimizer	Adam
Learning rate	2×10^{-3}
Weight decay	10^{-4}
LR schedule	cosine annealing, $T_{\max} = 120$
Max epochs	120
Early-stopping patience	16 (inner-val horizon AUROC)
Batch size	512
Inner-val split	0.85 / 0.15 (train fold only)
Ensemble size	3 (matched to CNN)
Cross-validation folds	5 (patient-level)
Seeds	5
SSL objective	SimCLR / NT-Xent on ~ 30 k unlabeled tensors
Spectrogram tensor	$F \approx 21 \times T = 40 \times L = 12$

For reproducibility, every model is trained under a fixed seed protocol: the five outer seeds index both the cross-validation shuffle and, within each fold, the three ensemble members (member s of seed g uses seed $10g+s$), with NumPy and PyTorch seeded identically and the global `RANDOM_STATE` fixed for cohort sampling and fold assignment. The pipeline is implemented in Python with PyTorch for the network and Cox loss, scikit-learn for the AUROC computation, SciPy for the short-time Fourier transform and filtering, and PyWavelets for wavelet denoising; device selection follows CUDA over Apple MPS over CPU. Full runs were executed on a SLURM cluster with NVIDIA A40 and A100 GPUs, with the multi-horizon and clinical-utility read-outs computed on CPU from saved per-patient risk scores. The cohort-construction, tensorization, model, and evaluation code is publicly available so that the gold ICD-10 results are reproducible end to end.

3.6. *Baselines*

Three baseline families are compared under a matched protocol. The clinical baseline uses age, sex, and ECG-interval measurements (P-wave duration, QRS duration, QT, and axes) in a Cox or logistic model, representing the risk-factor approach in the CHARGE-AF tradition [8]. The CNN baseline is a compact convolutional network on the same spectral tensor, serving as the black-box comparator, and uses a three-model ensemble to match BTTN. The block-term regression baseline is the real vendored linear block-term regressor [4], in both its standard and survival forms, which shows why a nonlinear generalization is needed.

3.7. *Evaluation*

All comparisons use patient-level cross-validation so that no patient appears in both training and test folds. A single primary comparison is pre-specified—conv-BTTN with self-supervised pretraining against the matched black-box CNN at the one-year horizon—and the architecture ladder, the block-count (K) sweep, and the two- and five-year horizons are treated as secondary, exploratory read-outs. Only the primary comparison is treated inferentially, carrying the paired patient-bootstrap confidence interval reported in Section 4; the secondary read-outs are point estimates with confidence intervals for description. No family-wise correction is applied, since a single comparison is the only one tested inferentially and the rest are descriptive. Confidence intervals at 95% come from patient-level bootstrap resampling,

377 which resamples whole patients and so respects the dependence between a pa-
 378 tient’s records, and methods are compared with paired tests on matched folds
 379 and seeds to control for fold-to-fold variation. The paired patient-bootstrap
 380 difference between two models, with its confidence interval and the fraction
 381 of resamples in which one model exceeds the other, is the primary basis
 382 for the accuracy claims; this follows the spirit of correlated-ROC compari-
 383 son [25], a parametric test on correlated ROC areas, applied here through
 384 a patient-level bootstrap. Beyond discrimination, the evaluation covers cal-
 385 ibration (Brier score and reliability), clinical utility through decision-curve
 386 net benefit [26] and operating points at fixed flag rates, and risk stratifica-
 387 tion by decile. Interpretability faithfulness is tested by fitting a linear probe
 388 on the pure interpretable block factors and measuring the fraction of full-
 389 model AUROC it recovers; a high fraction indicates that the interpretable
 390 subspace carries the prediction, with little left to the nonlinear remainder of
 391 the network. Generalization is assessed on the external CODE-15 cohort [27],
 392 examining both AF-in-sinus-rhythm identification and whether the learned
 393 frequency-and-lead substrate reproduces.

394 4. Results

395 All MIMIC-IV numbers below use gold ICD-10 I48 labels, evaluated with
 396 patient-level cross-validation and patient-bootstrap confidence intervals.

397 4.1. Discrimination

398 The headline result is a discrimination ladder that separates two claims:
 399 the interpretable ECG substrate is competitive with the black-box CNN, and
 400 fusing that substrate with clinical risk factors significantly beats both the risk
 401 factors alone and the black box. Table 3 traces the ladder at the one-year
 402 horizon ($n = 9,275$, five-seed ensemble AUROC). The clinical baseline of age,
 403 sex, and ECG intervals reached 0.760 [0.749, 0.771]. The plain linear block-
 404 term regressor on the spectral tensor reached 0.597 [0.582, 0.612], confirming
 405 that a linear model is capacity-limited on this input. The matched black-
 406 box CNN reached 0.734 [0.722, 0.745]. The interpretable conv-BTTN on the
 407 ECG alone reached 0.736 [0.724, 0.748], edging the CNN, and self-supervised
 408 pretraining lifted it to 0.747 [0.734, 0.759]; both sit below the clinical baseline,
 409 so the ECG substrate on its own is competitive with the black box but does
 410 not reach the risk factors. Adding the transparent clinical term raised conv-

411 BTTN to 0.776 [0.765, 0.787], and the full model with pretraining and fusion
 412 reached 0.778 [0.767, 0.789].

Table 3: One-year discrimination on gold ICD-10 MIMIC-IV labels ($n = 9,275$, patient-level cross-validation, five-seed ensemble AUROC with patient-bootstrap 95% CI). The ECG-only BTTN rows are competitive with the black-box CNN but below the clinical baseline; the fused model (ECG substrate plus the transparent clinical term) beats both.

Model	AUROC@1 yr [95% CI]
Clinical (age/sex + ECG intervals)	0.760 [0.749, 0.771]
Block-term regression (spectral, linear)	0.597 [0.582, 0.612]
CNN, ensembled (black box)	0.734 [0.722, 0.745]
Conv-BTTN (ECG-only)	0.736 [0.724, 0.748]
Conv-BTTN + SSL (ECG-only)	0.747 [0.734, 0.759]
Conv-BTTN + clinical	0.776 [0.765, 0.787]
Conv-BTTN + SSL + clinical (full)	0.778 [0.767, 0.789]

413 The fused model beats the clinical baseline by a paired $+0.019$ [$+0.014$,
 414 $+0.024$], and it exceeds clinical in 1.00 of the patient bootstrap. Significance
 415 here is read from the paired bootstrap on the same patients, which is more
 416 powerful than comparing the two marginal per-row confidence intervals. The
 417 intervals in Table 3 overlap. The paired difference cancels the shared patient-
 418 level variance and excludes zero decisively. The ladder isolates where the
 419 accuracy comes from. Self-supervised pretraining is a small lift on the ECG
 420 substrate ($+0.011$, from 0.736 to 0.747). Clinical-feature fusion is the large
 421 lever ($+0.031$ on top of the pretrained ECG model, from 0.747 to 0.778).
 422 The ECG substrate is not redundant with the risk factors: the full model
 423 adds $+0.019$ over the clinical baseline built from the same age, sex, and
 424 ECG intervals, so the interpretable spectral evidence carries information the
 425 interval measurements alone do not. The interpretable ECG-only model
 426 is competitive with the black-box CNN (0.747 versus 0.734). The fused
 427 model’s accuracy gain comes from transparent clinical fusion, not from any
 428 superiority of the block-term head over the black box.

429 The pattern holds at a longer horizon. On the five-year binary cohort
 430 ($n = 20,000$), the full model reached 0.794, an improvement of $+0.0223$
 431 [$+0.0187$, $+0.0261$] over the clinical baseline of 0.775 (probability of exceeding
 432 clinical 1.00), with the ECG-only model near 0.76.

4.2. Sensitivity to block count

The number of blocks K is the main structural hyperparameter, and BTTN is insensitive to it. Sweeping K from 4 to 24 moved AUROC by less than the width of its confidence interval, with the eight-to-twelve range a mild sweet spot and even four blocks sufficient. The headline result is therefore not the product of a cherry-picked block count.

4.3. Multi-horizon survival

A single survival model gives stable discrimination across horizons, a capability that a per-horizon binary classifier lacks. Reading the same fused survival model at 90 days and one, two, and five years without any per-horizon retraining, the BTTN-plus-clinical model reached AUROC 0.723, 0.727, 0.740, and 0.745, against 0.710, 0.711, 0.727, and 0.731 for the clinical baseline (Table 4). The concordance index over the full follow-up was 0.725 for the fused model, 0.714 for the clinical baseline, and 0.664 for the ECG substrate alone. The 90-day read-out is a true short-onset incident-*prediction* construct: positives are patients whose first AF falls in the window from index plus 30 days to index plus 120 days, so the 30-day blanking period is preserved in full, distinct from the near-term concurrent/identification task of Attia et al. [1]. These survival-model AUROCs are not comparable in absolute value to the binary Table 3 values: Table 3 reports a five-seed binary ensemble at a fixed horizon, whereas these come from one survival model read at several horizons. The table shows the cross-horizon capability and the concordance advantage; these are secondary to the headline discrimination number. The near-flat profile across horizons reinforces, as Section 4.6 shows, the reading of a stable, long-standing atrial substrate detectable from short to long horizons.

4.4. Clinical utility

Discrimination alone understates clinical value, so we examined calibration, stratification, and decision-analytic net benefit for the fused model (Fig. 3), with risk scores recalibrated to the natural one-year incidence of 4.25%. These read-outs use the survival-model risk (AUROC 0.726), which is distinct from the binary five-seed 0.778 headline. Risk stratification was strong: the top decile carried a one-year AF rate 20.7 times that of the bottom decile. Calibration was good, with a Brier score of 0.039 and a near-diagonal reliability curve. Decision-curve analysis placed the fused model above the clinical model and above the treat-all and treat-none defaults across the relevant threshold range. At a top-5% flag rate the fused model reached

Table 4: Multi-horizon discrimination from one gold-label survival model, read at each horizon with no per-horizon retraining. These single-survival-model values are not comparable in absolute magnitude to the five-seed binary ensemble of Table 3; they show the cross-horizon capability and the concordance advantage. ECG-only C-index 0.664.

Horizon	BTTN + clinical	Clinical
90-day onset	0.723	0.710
1 year	0.727	0.711
2 years	0.740	0.727
5 years	0.745	0.731
C-index	0.725	0.714

469 sensitivity 0.15, specificity 0.96, and positive predictive value 0.131, corre-
 470 sponding to a number-needed-to-monitor of 7.7 to detect one incident AF
 471 case; the number-needed-to-monitor was 8.7 at a top-10% flag rate and 10.2
 472 at top-20%, and the fused model beat the clinical baseline at every operat-
 473 ing point. Framed as enrichment for monitoring yield, a higher PPV in the
 474 flagged stratum lowers the number of patients who must be monitored to
 475 detect one incident AF case relative to the unselected population-screening
 476 approach of prospective AF-screening trials such as STROKESTOP [28], the
 477 wearable-patch mSToPS trial [29], and the implantable-loop-recorder LOOP
 478 trial [30], all of which monitor broadly selected populations; a risk model that
 479 concentrates monitoring on a high-yield stratum provides that enrichment.

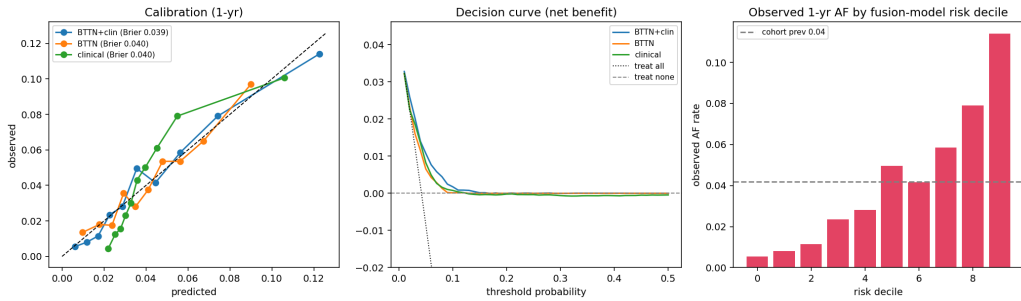


Figure 3: Clinical evaluation of the fused model on the one-year survival cohort, recalibrated to the natural 4.25% incidence: calibration reliability, decision-curve net benefit relative to clinical, treat-all, and treat-none strategies, and risk strata showing the 20.7-fold gradient between the top and bottom deciles.

480 4.5. Faithfulness of the interpretable factors

481 The central interpretability question is how much of BTTN’s decision
 482 its interpretable factors account for. Three complementary tests measure
 483 this directly: a linear probe with controls, an intervention test, and a block-
 484 structure ablation. Read together, they strongly support that a large part of
 485 the predictive signal is present in the interpretable representation, with one
 486 caveat we state up front: the linear-probe recovery is high but is not on its
 487 own sufficient, because a label-blind block layer recovers nearly as much, so
 488 the decisive evidence is the intervention test on the trained model rather than
 489 the recovery number. All three tests are computed on the ECG substrate,
 490 which fusion leaves untouched, under the same five-fold patient-level protocol
 491 with three seeds per fold; the full-model AUROC for this run is 0.718.

492 *Probe recovery with controls.* A linear readout fit on the pure interpretable
 493 block factors reached an AUROC of 0.701, against 0.718 for the full model,
 494 recovering 97.6% of the model’s accuracy. The probe input is the static,
 495 head-independent per-block scores: the time-averaged multilinear projection
 496 of the input spectrogram tensor onto the per-block frequency factors A_k
 497 and lead factors C_k , taken before the temporal-convolution, attention, and
 498 MLP head. These scores are exactly the head-independent quantities the
 499 interpretable factors define, so the probe measures how much of the decision
 500 lives in that subspace and cannot smuggle in head capacity. A recovery
 501 number is interpretable only against controls [14], so we fit the identical linear
 502 readout on four reference feature spaces (Table 5). The factor probe recovers
 503 0.976 of full-model accuracy, clearly above a probe on the raw flattened
 504 spectral tensor (0.870) and a matched-dimension random projection of the
 505 input (0.886), so the recovery reflects genuinely decodable structure in the
 506 learned subspace, beyond the near-linearity of the task itself. The black-
 507 box CNN’s penultimate features form the non-interpretable upper bound,
 508 which a linear probe recovers slightly past the full BTTN value (1.016). One
 509 nuance qualifies this. A label-blind block layer, whose factors never saw the
 510 labels, recovers nearly as much (0.973), so the recovery *number alone* does
 511 not isolate label-driven training; it largely reflects the readable projection
 512 geometry that any block layer imposes on the spectral input. The recovery
 513 number therefore establishes that the interpretable subspace is decodable,
 514 but it is the intervention test below, run on the trained model, that is the
 515 decisive proof that the decision is actually routed through the top-ranked
 516 interpretable factors.

517 *Intervention / sufficiency.* The decisive evidence that the highlighted

Table 5: Control probes for the faithfulness recovery (same linear logistic readout, patient-level cross-validation; recovery is probe AUROC divided by the full-model AUROC of 0.718). Three seeds, five folds.

Probe feature space	Recovery
BTTN interpretable factors	0.976
Raw spectral tensor	0.870
Random projection (matched dim)	0.886
Label-blind block factors	0.973
CNN penultimate (non-interp. ceiling)	1.016

518 factors are causal for the prediction is an intervention test on the trained
 519 model. The interpretable input axes (frequency bins and leads) were ranked
 520 by single-axis occlusion importance on a held-out fold (base AUROC 0.705),
 521 then progressively zeroed while recomputing full-model AUROC. The inter-
 522 vention conclusions are within-curve relative changes against this fold’s own
 523 baseline; they make no new absolute-performance claim. Ablating the top-
 524 ranked factors collapses the prediction monotonically, with AUROC falling
 525 by 0.014 when the single top factor is removed, by 0.049 at the top five,
 526 and by 0.086 at the top ten, whereas ablating an equal number of bottom-
 527 ranked factors leaves AUROC essentially unchanged. This asymmetry shows
 528 that the trained decision is routed specifically through the top-ranked inter-
 529 pretable factors, the model-internal sufficiency statement a recovery number
 530 alone cannot make.

531 *Block-structure ablation.* To test whether the block-term parameteriza-
 532 tion of the interpretable layer is doing work or is mere dimension reduction,
 533 we held the entire downstream network identical and swapped only the first
 534 interpretable layer among the block-term form (K blocks, each a learnable
 535 rank- $(r_f=3, r_l=2)$ frequency/lead subspace), a CP rank-1-per-block variant,
 536 and a single full-rank $F \times L$ projection per time step (Table 6). All three
 537 tie on accuracy (block-term 0.729, full-rank 0.729, CP rank-1 0.728), and
 538 block-term matches full-rank on faithfulness (0.959 versus 0.966) while the
 539 CP variant is a little lower (0.926). Block-term structure is therefore not
 540 uniquely required for accuracy, but it is the parameter-efficient interpretable
 541 instantiation: the low-rank per-block factors are exactly what yield the com-
 542 pact per-block frequency and lead signatures the interpretation in Section 4.6
 543 reads, at no accuracy cost relative to a dense projection. The result is in-

544 interpretability at the component level, established by holding accuracy fixed
 545 while varying only the interpretable layer.

Table 6: Block-structure ablation: only the interpretable first layer is swapped; the downstream conv head, Cox loss, and protocol are identical (five folds, three seeds). Faithfulness is the linear-probe recovery of full AUROC.

Interpretable layer	Full AUROC	Faithfulness
Block-term ($r_f=3, r_l=2$)	0.729	0.959
CP rank-1/block	0.728	0.926
Full-rank $F \times L$	0.729	0.966

546 The three tests support a quantitative, model-internal faithfulness state-
 547 ment: the interpretable subspace is decodable above non-interpretable con-
 548 trols, the trained decision is causally routed through its top-ranked factors,
 549 and the compact block-term parameterization delivers this readable struc-
 550 ture without sacrificing accuracy. The pattern argues against the concern
 551 that an interpretable bottleneck merely launders information leaked through
 552 a complex head.

553 4.6. Interpretability: substrate and cardiac-phase saliency

554 The learned substrate is physiologically coherent and recurs across read-
 555 outs (Fig. 4). Occluding individual frequency bands and leads in held-out
 556 inputs localized future-AF risk to the P-wave band near 2 Hz and a low-
 557 frequency atrial band at 4 to 8 Hz, with the most informative leads being I
 558 and II, the classic P-wave leads, together with V1 and V5. The term “fibril-
 559 latory band” is deliberately avoided for these sinus-rhythm recordings: the
 560 inputs contain no AF, so 4–8 Hz content here is not literal f-wave activity but
 561 low-frequency atrial-depolarization energy that overlaps the spectral range
 562 in which f-waves later appear, best read as a marker of pre-existing atrial
 563 substrate. These regions match long-established markers of atrial disease, in-
 564 cluding P-wave indices [31], advanced interatrial block (“Bayés syndrome”), a
 565 sinus-ECG conduction abnormality prospectively linked to incident AF and
 566 stroke [32], and the broader notion of atrial cardiopathy, the structural and
 567 electrical atrial substrate that precedes and predicts AF and thromboem-
 568 bolism independently of documented arrhythmia [33]. That the transparent
 569 occlusion map centers on the P-wave leads independently corroborates the
 570 black-box saliency analysis of Jabbour et al. [3], which found P-wave area

the most important feature of their deep model, reached here from an interpretable method with independent provenance. The substrate is confirmatory. That the model reads exactly this substrate from a sinus-rhythm ECG is consistent with the atrial-cardiopathy view that the disease is present before the first detected episode.

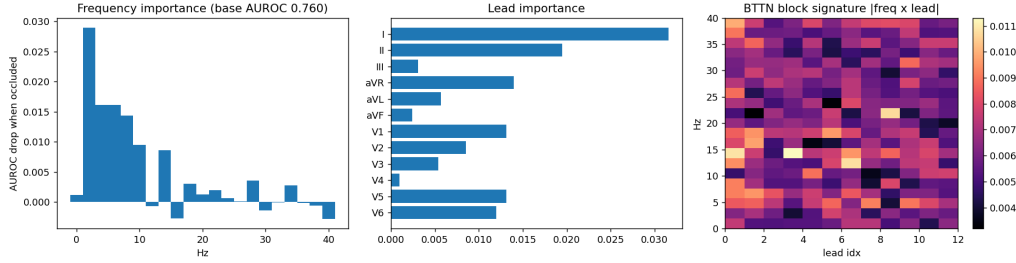


Figure 4: Glass-box substrate map on held-out MIMIC-IV data from band-and-lead occlusion. Future-AF risk localizes to the P-wave (~ 2 Hz) and low-frequency atrial (4–8 Hz, not literal f-waves in these sinus-rhythm inputs) bands and to leads I, II, V1, and V5.

A second analysis locates where in the cardiac cycle the signal lies, occluding fine windows aligned to the R-peak and rebuilding the spectrogram. Raw occlusion is dominated by the high-energy QRS complex, an artifact of removing the loudest waveform component, so the reported quantity is the energy-normalized saliency, the risk drop per unit of removed energy. Normalized this way, the atrial regions carry the most signal per unit amplitude, with the P-wave and PR phases ranking above QRS and ST/T (Fig. 5). The margins are small, so this ordering is directionally consistent with the frequency-domain substrate map and best read qualitatively. Only the energy-normalized version is reported, because raw occlusion saliency on the ECG is energy-confounded. A longitudinal version of this analysis, organized as cardiac phase against time before onset, shows the P-wave and PR atrial phases dominating throughout and remaining stable across the roughly three years before onset rather than emerging near it. This stability is coherent with the flat multi-horizon AUROC and supports a persistent, long-standing atrial substrate. The earlier appearance of a sharply rising peri-onset trajectory did not replicate at scale and is not claimed.

4.7. External validation

Generalization was tested on CODE-15 [27], an independent Brazilian cohort with a different population, device, and acquisition setting. Because

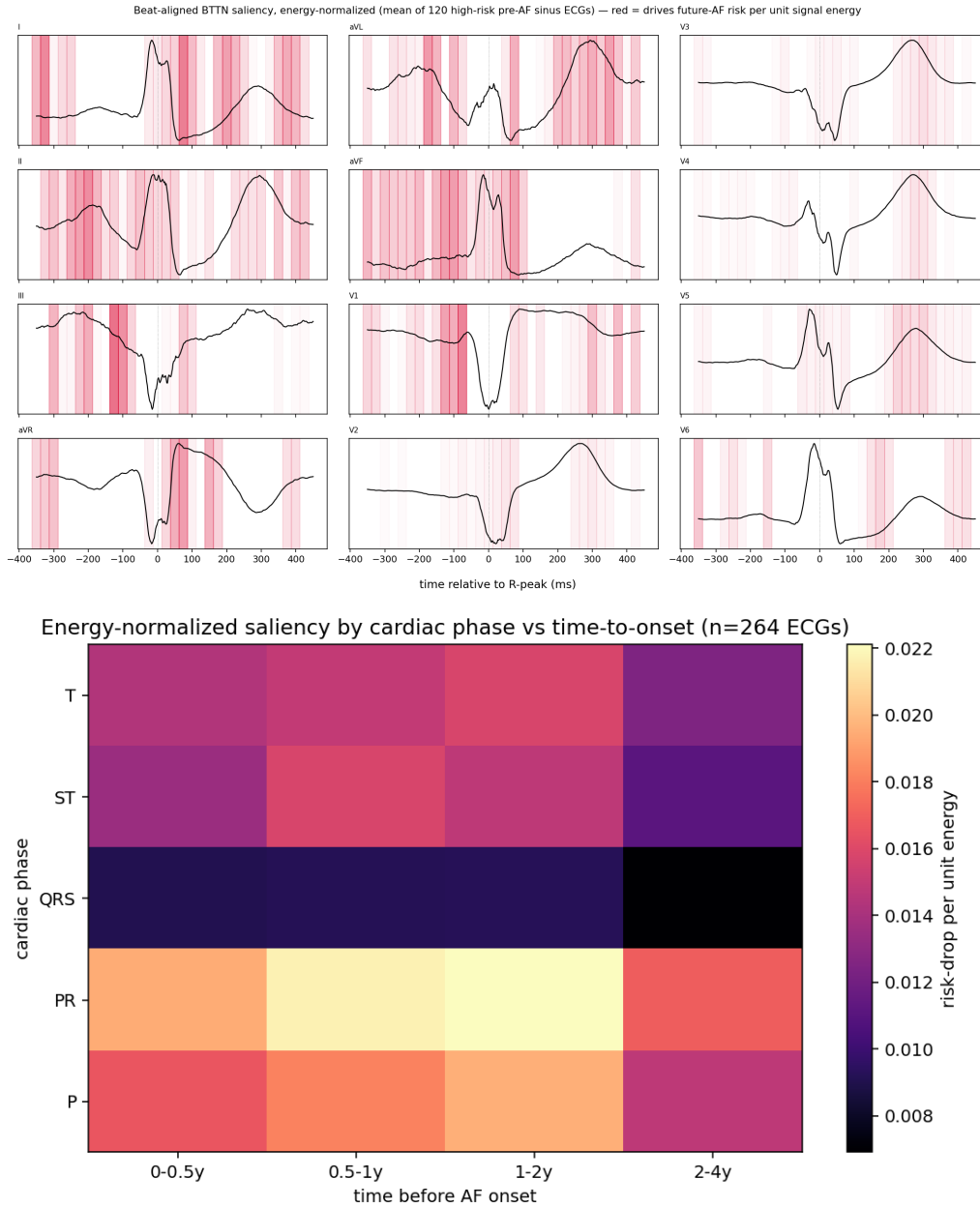


Figure 5: Energy-normalized cardiac-phase saliency. Top: risk drop per unit removed energy by cardiac phase, averaged over high-risk pre-AF ECGs, showing P-wave and PR phases above QRS and ST/T. Bottom: cardiac phase against time before onset, showing the atrial phases dominating and remaining stable across the years preceding onset.

CODE-15 lacks dated follow-up, this section evaluates *substrate transfer* together with two cross-sectional tasks—AF-in-sinus-rhythm *identification* and SR-versus-AF *detection*—both of which are distinct from, and easier than, the dated future-onset prediction task that is the subject of this paper; none of the AUROCs below speak to the incidence claim. The primary external read-out is therefore substrate reproduction. The learned substrate reproduced: V1 was the top lead in both cohorts, with I and II also shared, and the highlighted frequencies fell in the low-frequency atrial band (6.7 to 10 Hz on CODE-15, against 4–8 Hz on MIMIC-IV) and the low-frequency P-wave band (3.3 Hz, against 2 Hz on MIMIC-IV), with importances large enough to exclude noise (Fig. 6). As supporting evidence that the pipeline is not specific to MIMIC-IV, the AF-in-sinus-rhythm *identification* task (n = 7,968, 1,968 positive) gave an AUROC of 0.829, the SR-versus-AF *detection* task (n = 10,799, 4,799 positive) gave 0.980 on CODE-15, and the same spectral pipeline reached 0.983 for SR-versus-AF *detection* (n = 3,140, 1,570 positive) on PTB-XL [34], an independent dataset and device. These detection and identification AUROCs are not incidence-prediction performance; they corroborate substrate transfer only. These results indicate that the interpretable substrate is a property of the physiology the model learns and holds across datasets.

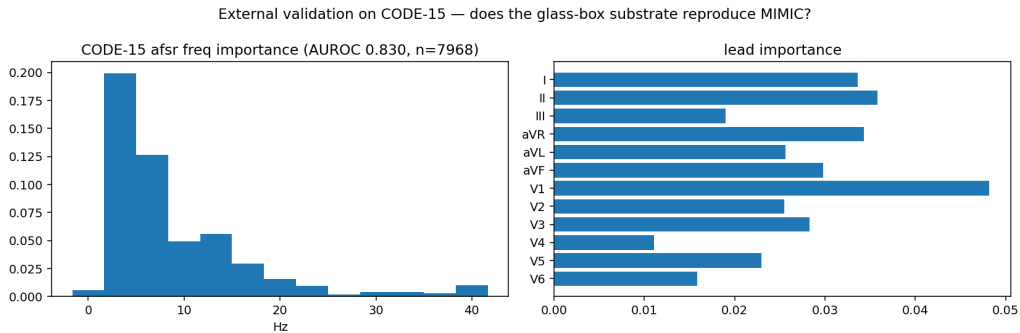


Figure 6: External validation on CODE-15: substrate transfer is the primary read-out (top lead V1; low-frequency atrial band 6.7–10 Hz; P-wave band 3.3 Hz), consistent with the MIMIC-IV maps. The AF-in-sinus-rhythm *identification* AUROC of 0.829 (n = 7,968, 1,968 positive) is a different and easier task than dated incidence and is shown as supporting evidence only, not as incidence-prediction performance.

616 5. Discussion

617 The distinctive contribution is a glass-box model whose interpretable fac-
618 tors are shown, by direct measurement rather than post-hoc assertion, to ac-
619 count for a large part of its decision. Accuracy is competitive rather than the
620 differentiator: the transparent fusion of the ECG substrate with clinical risk
621 factors significantly outperforms both a risk-factor baseline and a matched
622 black-box CNN, while every prediction rests on factor-level evidence. On the
623 gold ICD-10 cohort, the interpretable ECG-only model was competitive with
624 the CNN (0.747 versus 0.734) and below the clinical baseline of age, sex, and
625 ECG intervals (0.760); the two sources of information are complementary
626 rather than redundant. Fusing them through a zero-initialized additive term
627 that bypasses the block layer raised the one-year AUROC to 0.778, a paired
628 +0.019 over the clinical baseline (probability of exceeding clinical 1.00), and
629 beat the CNN. The fusion term never touches the interpretable projection, so
630 this gain comes at no interpretability cost: the per-block frequency and lead
631 signatures and the faithfulness probe are identical to the ECG-only model.
632 The claim is interpretability at competitive accuracy through transparent
633 fusion, not a method-superiority accuracy record over the black box.

634 The progression from the linear baseline is informative about why the in-
635 terpretable structure needed to change. A linear block-term regressor on the
636 spectral tensor reached only 0.597, well short of the CNN, which confirms that
637 the interpretable structure alone cannot carry this task. Each subsequent
638 step was an architectural change rather than a tuning choice: a nonlinear
639 head, a multi-scale temporal convolution that replaces the static block score
640 with a learnable local operator, ensembling, self-supervised pretraining, and
641 the additive clinical term. The interpretable factors were retained through
642 the entire progression. Capacity was added around them while keeping them
643 in place, so a linear probe on those factors still recovers almost all of the
644 ECG-substrate accuracy.

645 The accuracy should be read against the field, and read honestly. The
646 strongest one-year incident-AF model, Raghunath et al. [2], reported AU-
647 ROC near 0.85 on a roughly 430,000-patient proprietary Geisinger cohort;
648 the one-year fused model here reaches 0.778, below that record. At five
649 years, the fused model reached about 0.79, against the ECG-AI model of
650 Jabbour et al. [3], which reported 0.77 to 0.78 for five-year incident AF
651 on MIMIC-IV. This is, to our knowledge, the highest AUROC reported for
652 MIMIC-IV five-year incident-AF prediction, with one caveat: the figure here

653 is in-distribution, trained and tested within MIMIC-IV, whereas the 0.77 of
654 Jabbour et al. is external transfer validation of a model trained on the Mon-
655 treal Heart Institute cohort, and the two studies differ in cohort and label
656 construction. The result is therefore the strongest we are aware of on this
657 dataset, not a demonstrated method-superiority over a black box. Jabbour et
658 al. found that adding CHARGE-AF clinical risk factors and polygenic scores
659 did not improve their ECG-AI model, whereas the smaller in-distribution
660 model here gains materially from explicit clinical fusion, consistent with the
661 ECG and the risk factors carrying complementary information at this scale.

662 The faithfulness measurement is what distinguishes this interpretability
663 claim from the usual post-hoc account. A linear probe on the pure block
664 factors recovered 97.6% of the full model’s accuracy, above a probe on the
665 raw spectral tensor or a matched random projection, and an intervention
666 test showed that ablating the top-ranked interpretable factors collapses the
667 trained model while ablating the bottom-ranked ones leaves it intact. Be-
668 cause a label-blind block layer recovers nearly as much under the probe alone,
669 the recovery figure is read as strong support that a large part of the predic-
670 tive signal lives in the interpretable subspace, and the intervention test on
671 the trained model is the load-bearing evidence that the decision is routed
672 through the top-ranked factors. The measurement directly addresses the
673 concept-bottleneck-leakage concern [16], under which an interpretable layer
674 can appear meaningful while the real work happens in an opaque head. A
675 formal no-leakage guarantee is out of scope, and by-design multilinear inter-
676 pretability has been pursued elsewhere [17]. The contribution is the mea-
677 surement itself, applied to a spectral-ECG survival model.

678 The survival formulation adds a clinically useful property. The model
679 is trained on continuous time-to-event with a Cox loss, so one model serves
680 all horizons, with AUROC essentially flat from 90 days to five years and a
681 concordance index of 0.725 for the fused model against 0.714 for the clinical
682 baseline and 0.664 for the ECG substrate alone. A per-horizon binary clas-
683 sifier would require separate models and could not express the smooth risk
684 profile that the survival output provides. The flat horizon curve is consis-
685 tent with the interpretability read-outs, the longitudinal substrate trajectory
686 and the cardiac-phase saliency, which point to a persistent atrial substrate
687 present years before onset.

688 The substrate the model learns is clinically coherent and reproducible.
689 Low-frequency P-wave content, the low-frequency atrial 4–8 Hz band (low-
690 frequency atrial-depolarization energy in these sinus-rhythm inputs), and the

691 P-wave leads I, II, and V1 are established correlates of atrial disease, includ-
 692 ing the P-wave indices of [31], and BTTN recovers them on MIMIC-IV and
 693 again on CODE-15 across a different population and device. The transparent
 694 occlusion map lands on the same P-wave features that the black-box saliency
 695 analysis of Jabbour et al. [3] flagged as most important, reached here from an
 696 interpretable method with independent provenance. The cardiac-phase anal-
 697 ysis, once corrected for the energy bias that otherwise inflates QRS, places
 698 the signal in the P-wave and PR atrial phases. All of this is confirmatory.
 699 The substrate is already established; the finding is that a competitive pre-
 700 dictive model recovers it transparently and that the same substrate survives
 701 external transfer. From a clinical-utility standpoint, the 20.7-fold relative
 702 risk gradient between the top and bottom deciles, the favorable decision-
 703 curve net benefit, and the operating points recalibrated to the natural 4.25%
 704 incidence (number-needed-to-monitor 7.7 at a top-5% flag rate) suggest that
 705 the fused model could enrich monitoring yield by concentrating rhythm mon-
 706 itoring on a higher-risk stratum. The read-out is enrichment for monitoring
 707 yield, not a clinically actionable decision rule.

708 The energy-bias correction in the cardiac-phase analysis deserves empha-
 709 sis because it is a general pitfall for occlusion-based saliency on physiological
 710 signals. Removing a high-amplitude component such as the QRS complex
 711 perturbs a magnitude spectrogram far more than removing a low-amplitude
 712 component such as the P-wave, so an uncorrected occlusion map will report
 713 QRS as most important regardless of where the discriminative information
 714 lies. Normalizing the risk drop by the energy removed restores the ordering
 715 that the frequency-domain substrate map already implied, with the atrial
 716 phases carrying the most signal per unit amplitude. Only the normalized
 717 version is reported for this reason, and the raw map is flagged as confounded.
 718 The same caution applies to any future attempt to localize ECG-model de-
 719 cisions in the time domain.

720 Several design choices were deliberately kept lean. The block count does
 721 not need tuning, since accuracy is flat from four to twenty-four blocks, and ex-
 722 tending the nonlinear part of the network beyond the multi-scale convolution
 723 did not improve accuracy in our experiments. A lean architecture is simpler
 724 to train and keeps the interpretability claim tight: the more the prediction
 725 depends on an opaque downstream stack, the less the interpretable factors
 726 can explain. The measured faithfulness of 97.6% is consistent with this: the
 727 interpretable subspace carries the prediction precisely because the network
 728 does not lean on a deep, opaque head. Self-supervised pretraining provided a

small, consistent gain on the ECG substrate, and the clinical-feature fusion, added through a transparent linear term, is the larger accuracy lever.

6. Limitations

Several limitations bound these results. First, the development cohort is a single internal source. The external CODE-15 and PTB-XL analyses validate AF identification and substrate transfer, not dated incidence, because neither external cohort carries the dated future-onset outcome the prediction task requires, so external validation of the incidence claim itself remains open. Second, positioning against the field is honest but limited. Our one-year fused model (0.778) is below the roughly 0.85 one-year record of Raghunath et al. [2], obtained on a far larger proprietary cohort, and while our five-year figure is the strongest we are aware of on MIMIC-IV, it is in-distribution rather than the external transfer validation reported by Jabbour et al. [3], so it is not a demonstrated method-superiority over a black box. Third, a competing-risks analysis is not yet activated: it requires date-of-death information to correct the upward bias in AF risk when a patient dies of a competing cause before AF, and the gold-ready variant of Appendix Appendix A awaits that field. Fourth, we deliberately excluded a near-term (30 to 31 day) onset analysis, because at such short lead times gold-ICD ascertainment conflates prediction with detection of AF already emerging, which is a different task and admits no fair comparison with the near-concurrent identification of Attia et al. [1]. Fifth, the interpretability contribution is confirmatory rather than novel. The attribution tools are standard, the P-wave and V1 atrial substrate is already established and externally validated in prior work, and by-design multilinear interpretability has been explored elsewhere; our defensible contribution is the measured faithfulness of the model’s own factors, not a new interpretability method or physiological discovery. Finally, the cardiac-phase saliency margins are small and the longitudinal per-patient trajectories are noisy, so those read-outs are best regarded as directionally supportive, with the numbers too noisy to serve as precise effect estimates.

7. Conclusion

We presented BTTN, a learnable nonlinear survival generalization of block-term tensor regression for incident-AF prediction from the sinus-rhythm

12-lead ECG, fused transparently with clinical risk factors. On a gold ICD-10 cohort, the interpretable ECG substrate was competitive with a matched black-box CNN, and fusing it with age, sex, and ECG intervals through a zero-initialized additive term that bypasses the interpretable layer significantly beat both the risk-factor baseline and the black box, at no interpretability cost. The model delivered stable multi-horizon risk from a single survival model and offered enrichment for monitoring yield with a 20.7-fold risk gradient and favorable net benefit recalibrated to the natural incidence. Its interpretable frequency and lead factors recovered 97.6% of full-model accuracy in a linear probe and routed the trained decision through the top-ranked factors under an intervention test, making the interpretability a measured property, and the learned atrial substrate reproduced on an independent external cohort. The immediate next step is to secure an external cohort with dated outcomes for direct validation of the incidence claim and to activate the competing-risks model with date-of-death information.

Data Availability

MIMIC-IV-ECG [23], PTB-XL [34], and CODE-15 [27] are publicly available through their respective repositories under their data-use agreements. The MIMIC-IV hospital module required for gold ICD-10 labels is governed by a credentialed PhysioNet agreement. Code to reproduce the cohort construction, tensorization, model, and evaluation is publicly available at <https://github.com/TheAxeC/bttn-incident-af>.

Ethics

This study uses de-identified, publicly available datasets collected under the governance of their originating institutions and made available through credentialed data-use agreements. No new human-subjects data were collected, and the analysis was conducted in accordance with the terms of each dataset’s agreement.

CRedit Author Statement

Axel Faes: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing (original draft). **Liesbet Peeters:** Supervision, Resources, Funding acquisition, Writing (review and editing).

794 **Conflicts of Interest**

795 The authors declare no competing financial or non-financial interests.

796 **Funding**

797 This work was supported by the Flanders AI Research Program (FAIR),
798 an Impulse Program of the Flemish Government (Flemish Ministry of Econ-
799 omy, Science and Innovation).

800 **AI-Usage Disclosure**

801 AI-based writing-assistance tools were used to help draft and edit the
802 manuscript text. All scientific content, analyses, numerical results, and in-
803 terpretations were produced and verified by the authors, who take full re-
804 sponsibility for the work.

805 **Appendix A. Competing-risks survival variant (gold-ready)**

806 The competing-risks variant referenced in Section 3 is a discrete-time,
807 cause-specific deep survival model. The time axis is binned, and the head
808 emits a softmax over the joint space of cause and bin for the two causes AF
809 and death. The loss combines a negative log-likelihood term, which places
810 probability mass in an event’s own cause-and-bin cell or in the survival region
811 beyond a censored patient’s bin, with a cause-specific ranking term that
812 encourages correct ordering of AF risk. The intent is to remove the upward
813 bias in AF risk that naive right-censoring introduces when a patient dies of
814 a competing cause before AF could occur. The variant currently passes an
815 implementation sanity check on synthetic competing-risks data (AF-cause
816 C-index 0.867); it activates once gold date-of-death information is available
817 and has not been run on the present cohort.

818 **References**

- 819 [1] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J.
820 Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J.
821 Erickson, S. Kapa, P. A. Friedman, An artificial intelligence-enabled
822 ECG algorithm for the identification of patients with atrial fibrillation
823 during sinus rhythm: a retrospective analysis of outcome prediction, The

- 824 Lancet 394 (10201) (2019) 861–867. doi:10.1016/S0140-6736(19)31721-
825 0.
- 826 [2] S. Raghunath, J. M. Pfeifer, A. E. Ulloa-Cerna, A. Nemani, T. Carbon-
827 ati, L. Jing, D. P. vanMaanen, D. N. Hartzel, J. A. Ruhl, B. F. Lager-
828 man, D. B. Rocha, N. J. Stoudt, G. Schneider, K. W. Johnson, N. Zim-
829 merman, J. B. Leader, H. L. Kirchner, C. J. Griessenauer, A. Hafez,
830 C. W. Good, B. K. Fornwalt, C. M. Haggerty, Deep neural networks can
831 predict new-onset atrial fibrillation from the 12-lead ECG and help iden-
832 tify those at risk of atrial fibrillation-related stroke, *Circulation* 143 (13)
833 (2021) 1287–1298. doi:10.1161/CIRCULATIONAHA.120.047829.
- 834 [3] G. Jabbour, A. Nolin-Lapalme, O. Tastet, D. Corbin, P. Jordà, A. Sowa,
835 J. Delfrate, D. Busseuil, J. G. Hussin, M.-P. Dubé, J.-C. Tardif, L. Ri-
836 vard, L. Macle, J. Cadrin-Tourigny, P. Khairy, R. Avram, R. Tadros,
837 Prediction of incident atrial fibrillation using deep learning, clinical mod-
838 els, and polygenic scores, *European Heart Journal* 45 (46) (2024) 4920–
839 4934. doi:10.1093/eurheartj/ehae595.
- 840 [4] A. Faes, F. Camarrone, M. M. Van Hulle, Single finger trajectory pre-
841 diction from intracranial brain activity using block-term tensor regres-
842 sion with fast and automatic component extraction, *IEEE Transactions*
843 *on Neural Networks and Learning Systems* 35 (7) (2024) 8897–8908.
844 doi:10.1109/TNNLS.2022.3216589.
- 845 [5] G. Christopoulos, J. Graff-Radford, C. L. Lopez, X. Yao, Z. I. At-
846 tia, A. A. Rabinstein, R. C. Petersen, D. S. Knopman, M. M. Mielke,
847 W. Kremers, P. Vemuri, K. C. Siontis, P. A. Friedman, P. A. Noseworthy,
848 Artificial intelligence–electrocardiography to predict incident atrial fib-
849 rillation, *Circulation: Arrhythmia and Electrophysiology* 13 (12) (2020)
850 e009355. doi:10.1161/CIRCEP.120.009355.
- 851 [6] M. Gadaleta, P. Harrington, E. Barnhill, E. Hytopoulos, M. P. Turakhia,
852 S. R. Steinhubl, G. Quer, Prediction of atrial fibrillation from at-home
853 single-lead ECG signals without arrhythmias, *npj Digital Medicine* 6 (1)
854 (2023) 229. doi:10.1038/s41746-023-00966-w.
- 855 [7] S. Khurshid, S. Friedman, C. Reeder, P. Di Achille, N. Diamant,
856 P. Singh, L. X. Harrington, X. Wang, M. A. Al-Alusi, G. Sarma, A. S.
857 Foulkes, P. T. Ellinor, C. D. Anderson, J. E. Ho, A. A. Philippakis,

- 858 P. Batra, S. A. Lubitz, ECG-based deep learning and clinical risk fac-
 859 tors to predict atrial fibrillation, *Circulation* 145 (2) (2022) 122–133.
 860 doi:10.1161/CIRCULATIONAHA.121.057480.
- 861 [8] A. Alonso, B. P. Krijthe, T. Aspelund, K. A. Stepas, M. J. Pencina,
 862 C. B. Moser, M. F. Sinner, N. Sotoodehnia, J. D. Fontes, A. C. J. W.
 863 Janssens, R. A. Kronmal, J. W. Magnani, J. C. Witteman, A. M.
 864 Chamberlain, S. A. Lubitz, R. B. Schnabel, S. K. Agarwal, D. D.
 865 McManus, P. T. Ellinor, M. G. Larson, G. L. Burke, L. J. Launer,
 866 A. Hofman, D. Levy, J. S. Gottdiener, S. Kääh, D. Couper, T. B. Har-
 867 ris, E. Z. Soliman, B. H. C. Stricker, V. Gudnason, S. R. Heckbert,
 868 E. J. Benjamin, Simple risk model predicts incidence of atrial fibrilla-
 869 tion in a racially and geographically diverse population: the CHARGE-
 870 AF consortium, *Journal of the American Heart Association* 2 (2) (2013)
 871 e000102. doi:10.1161/JAHA.112.000102.
- 872 [9] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, Y. Kluger,
 873 DeepSurv: personalized treatment recommender system using a Cox
 874 proportional hazards deep neural network, *BMC Medical Research*
 875 *Methodology* 18 (1) (2018) 24. doi:10.1186/s12874-018-0482-1.
- 876 [10] M. F. Gensheimer, B. Narasimhan, A scalable discrete-time survival
 877 model for neural networks, *PeerJ* 7 (2019) e6257. doi:10.7717/peerj.6257.
- 878 [11] C. Lee, W. R. Zame, J. Yoon, M. van der Schaar, DeepHit: A deep learn-
 879 ing approach to survival analysis with competing risks, in: *Proceedings*
 880 *of the AAAI Conference on Artificial Intelligence*, Vol. 32, 2018, pp.
 881 2314–2321. doi:10.1609/aaai.v32i1.11842.
- 882 [12] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim,
 883 Sanity checks for saliency maps, in: *Advances in Neural Information*
 884 *Processing Systems (NeurIPS)*, Vol. 31, 2018, pp. 9505–9515.
- 885 [13] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems:
 886 How should we define and evaluate faithfulness?, in: *Proceedings of the*
 887 *58th Annual Meeting of the Association for Computational Linguistics*
 888 *(ACL)*, 2020, pp. 4198–4205. doi:10.18653/v1/2020.acl-main.386.
- 889 [14] J. Hewitt, P. Liang, Designing and interpreting probes with control
 890 tasks, in: *Proceedings of the 2019 Conference on Empirical Methods*

- 891 in Natural Language Processing and the 9th International Joint Con-
 892 ference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp.
 893 2733–2743. doi:10.18653/v1/D19-1275.
- 894 [15] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim,
 895 P. Liang, Concept bottleneck models, in: Proceedings of the 37th Inter-
 896 national Conference on Machine Learning (ICML), Vol. 119 of Proceed-
 897 ings of Machine Learning Research, PMLR, 2020, pp. 5338–5348.
- 898 [16] M. Havasi, S. Parbhoo, F. Doshi-Velez, Addressing leakage in concept
 899 bottleneck models, in: Advances in Neural Information Processing Sys-
 900 tems (NeurIPS), Vol. 35, 2022, pp. 23386–23397.
- 901 [17] T. Dooms, W. Gauderis, G. A. Wiggins, J. Oramas, Composition-
 902 ality unlocks deep interpretable models, [https://arxiv.org/abs/2504.](https://arxiv.org/abs/2504.02667)
 903 02667, arXiv:2504.02667 (2025).
- 904 [18] T. G. Kolda, B. W. Bader, Tensor decompositions and applications,
 905 SIAM Review 51 (3) (2009) 455–500. doi:10.1137/07070111X.
- 906 [19] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa,
 907 H. A. Phan, Tensor decompositions for signal processing applications:
 908 From two-way to multiway component analysis, IEEE Signal Processing
 909 Magazine 32 (2) (2015) 145–163. doi:10.1109/MSP.2013.2297439.
- 910 [20] L. De Lathauwer, Decompositions of a higher-order tensor in
 911 block terms—Part II: definitions and uniqueness, SIAM Journal
 912 on Matrix Analysis and Applications 30 (3) (2008) 1033–1066.
 913 doi:10.1137/070690729.
- 914 [21] B. Hunyadi, D. Camps, L. Sorber, W. Van Paesschen, M. De Vos,
 915 S. Van Huffel, L. De Lathauwer, Block term decomposition for modelling
 916 epileptic seizures, EURASIP Journal on Advances in Signal Processing
 917 2014 (1) (2014) 139. doi:10.1186/1687-6180-2014-139.
- 918 [22] J. Ye, L. Wang, G. Li, D. Chen, S. Zhe, X. Chu, Z. Xu, Learn-
 919 ing compact recurrent neural networks with block-term tensor de-
 920 composition, in: Proceedings of the IEEE/CVF Conference on Com-
 921 puter Vision and Pattern Recognition (CVPR), 2018, pp. 9378–9387.
 922 doi:10.1109/CVPR.2018.00977.

- 923 [23] B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, C. Fernan-
 924 des, N. Greenbaum, J. W. Waks, P. Eslami, T. Carbonati, A. Chaud-
 925 hari, E. Herbst, D. Moukheiber, S. Berkowitz, R. Mark, S. Horng,
 926 MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset (version
 927 1.0), <https://physionet.org/content/mimic-iv-ecg/1.0/> (2023).
 928 doi:10.13026/4nqg-sb35.
- 929 [24] D. R. Cox, Regression models and life-tables, *Journal of the Royal*
 930 *Statistical Society: Series B (Methodological)* 34 (2) (1972) 187–202.
 931 doi:10.1111/j.2517-6161.1972.tb00899.x.
- 932 [25] E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the
 933 areas under two or more correlated receiver operating characteristic
 934 curves: a nonparametric approach, *Biometrics* 44 (3) (1988) 837–845.
 935 doi:10.2307/2531595.
- 936 [26] A. J. Vickers, E. B. Elkin, Decision curve analysis: a novel method for
 937 evaluating prediction models, *Medical Decision Making* 26 (6) (2006)
 938 565–574. doi:10.1177/0272989X06295361.
- 939 [27] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R.
 940 Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Mac-
 941 farlane, W. Meira Jr, T. B. Schön, A. L. P. Ribeiro, Automatic diagnosis
 942 of the 12-lead ECG using a deep neural network, *Nature Communica-*
 943 *tions* 11 (1) (2020) 1760. doi:10.1038/s41467-020-15432-4.
- 944 [28] E. Svennberg, L. Friberg, V. Frykman, F. Al-Khalili, J. Engdahl,
 945 M. Rosenqvist, Clinical outcomes in systematic screening for atrial fibril-
 946 lation (STROKESTOP): a multicentre, parallel group, unmasked, ran-
 947 domised controlled trial, *The Lancet* 398 (10310) (2021) 1498–1506.
 948 doi:10.1016/S0140-6736(21)01637-8.
- 949 [29] S. R. Steinhubl, J. Waalen, A. M. Edwards, L. M. Ariniello, R. R.
 950 Mehta, G. S. Ebner, C. Carter, K. Baca-Motes, E. Felicione, T. Sarich,
 951 E. J. Topol, Effect of a home-based wearable continuous ECG mon-
 952 itoring patch on detection of undiagnosed atrial fibrillation: the
 953 mSToPS randomized clinical trial, *JAMA* 320 (2) (2018) 146–155.
 954 doi:10.1001/jama.2018.8102.

- 955 [30] J. H. Svendsen, S. Z. Diederichsen, S. Højberg, D. W. Krieger, C. Graff,
956 C. Kronborg, M. S. Olesen, J. B. Nielsen, A. G. Holst, A. Brandes,
957 L. Køber, K. J. Haugan, Implantable loop recorder detection of atrial fibrillation to prevent stroke (the LOOP study): a randomised controlled
958 trial, *The Lancet* 398 (10310) (2021) 1507–1516. doi:10.1016/S0140-
959 6736(21)01698-6.
- 961 [31] J. W. Magnani, L. Zhu, F. Lopez, M. J. Pencina, S. K. Agarwal, E. Z.
962 Soliman, E. J. Benjamin, A. Alonso, P-wave indices and atrial fibrillation: cross-cohort assessments from the Framingham Heart Study (FHS)
963 and Atherosclerosis Risk in Communities (ARIC) study, *American Heart Journal* 169 (1) (2015) 53–61.e1. doi:10.1016/j.ahj.2014.10.009.
- 966 [32] M. Martínez-Sellés, R. Elosua, M. Ibarrola, M. de Andrés, P. Díez-
967 Villanueva, A. Bayés-Genis, A. Baranchuk, A. Bayés-de Luna, Advanced
968 interatrial block and P-wave duration are associated with atrial fibrillation and stroke in older adults with heart disease: the BAYES registry,
969 *EP Europace* 22 (7) (2020) 1001–1008. doi:10.1093/europace/euaa114.
- 971 [33] H. Kamel, P. M. Okin, W. T. Longstreth, M. S. V. Elkind, E. Z. Soli-
972 man, Atrial cardiopathy: a broadened concept of left atrial thromboem-
973 bolism beyond atrial fibrillation, *Future Cardiology* 11 (3) (2015) 323–
974 331. doi:10.2217/fca.15.22.
- 975 [34] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze,
976 W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocar-
977 diography dataset, *Scientific Data* 7 (1) (2020) 154. doi:10.1038/s41597-
978 020-0495-6.