Can We Trust Concept-Based ECG Explanations? A TCAV Analysis of Confounding and Reproducibility in Atrial Fibrillation Detection

El Hassane Jennah*, Houda Indjaren, Lhoucine Ben Taleb, Hamid El Malali, Azeddine Mouhsen
Department of Applied Physics, Energy, Matter, Instrumentation and Telecom Laboratory-Faculty of Sciences and Techniques, Hassan First University, Settat, Morocco

***Abstract*—AI explanation methods are typically validated in one of two ways: by confirming that an explanation is not an artifact of the underlying data (a confound check), or by confirming that it generalizes to unseen data. Fewer studies examine whether an explanation is stable across independently trained copies of the same model. We present a validation framework combining both dimensions — confound control and model-seed reproducibility — and demonstrate it in a case study using a deep learning model trained to detect atrial fibrillation (AF) from long, continuous ECG recordings. Using concept activation vectors (CAVs), we test whether the model's predictions are associated with three clinically meaningful patterns: irregular heartbeat timing, the presence of abnormal beats, and proximity to a rhythm change. We find that a concept-prediction association can be statistically significant and survive a targeted AF-label confound control in one trained model, yet reverse direction when the same architecture is retrained from a different random initialization. Across the nine concept-depth combinations examined, three replicated consistently across five independently trained models, two were inconclusive, and four showed evidence of directional instability under the predefined criterion. These findings demonstrate, within this AFNet/LTAFDB case study, that statistical significance and confound robustness alone may not establish a stable concept-based explanation; model-seed reproducibility provides a complementary validation dimension that can reveal sensitivity to training initialization.**

***Keywords*—Explainable AI; concept activation vectors; TCAV; atrial fibrillation; reproducibility; electrocardiogram**

I. Introduction

Deep learning models can detect atrial fibrillation (AF) from ECG signals with high accuracy [1]-[3]but their internal reasoning remains difficult to interpret. One approach to explaining what a model has learned is to test whether its internal representation aligns with a clinically meaningful concept (e.g., “irregular heartbeat”) rather than only identifying which input samples influenced a prediction. This is known as concept-based explanation, and Testing with Concept Activation Vectors (TCAV) is a widely used method of this kind [4]Subsequent work has extended it toward automatic concept discovery and completeness guarantees [5], [6]. Existing studies typically validate a result against a single trained model, sometimes after ruling out an obvious confound (for instance, confirming that the “concept” is not simply a proxy for the prediction itself). We ask a further question: if a result survives such a check, does it still hold when the same model is retrained from scratch?

We investigate this using AFNet, a compact convolutional network trained to detect AF from the Long Term AF Database, a set of continuous, day-long ambulatory ECG recordings, as a testbed for a more general validation framework. We define three clinical concepts directly from the dataset's own annotations — RR-interval irregularity, ectopic beat burden, and proximity to a rhythm change — rather than relying on externally curated example images, as in prior work. We test whether the model's predictions are associated with each concept at three depths within the network, both with and without an AF-label confound control, and across five independently trained models. The principal finding of this study concerns not any single clinical pattern, but a methodological one: a concept-prediction association can be statistically significant and pass a confound check, yet still fail to reproduce when the model is retrained from a different initialization — confound robustness and model-seed reproducibility are distinct validation dimensions that can disagree.

II. Related Work

Earlier work applying TCAV-style methods to ECG models treats a “concept” as a diagnostic label (for example, “this ECG shows AF”), built from a hand-picked set of example recordings, tested on short 10-second clips, and validated against only one trained model [7]. A similar approach has been used for EEG [8]. More broadly, XAI research has emphasized the need for explicit evaluation criteria rather than assuming that an explanation is valid because it is visually plausible or statistically associated with a target [9]-[11]. In particular, rigorous evaluation frameworks distinguish between different explanation properties and recognize that explanation behavior can depend on the underlying model and data [12]. Saliency-method work has also shown that model- and data-randomization tests can reveal explanations that are insufficiently tied to the learned model. In healthcare, recent surveys and systematic reviews likewise identify explanation assessment and reliability as important open issues [9], [13], [14]. Most other ECG and medical-imaging explanation work instead highlights which parts of the input mattered (e.g., saliency maps, SHAP, Grad-CAM), which represents a different form of explanation, often validated through visual or attribution-based analyses [15]-[16]. Our

*Corresponding author.

work differs in three respects: our concepts are operationalized directly from the dataset's own beat and rhythm annotations rather than hand-picked; we use continuous, multi-hour recordings rather than short clips; and we add a systematic reproducibility check across several independently trained models — reframing the contribution from "TCAV applied to ECG" toward a systematic confound-and-model-seed validation framework for concept-based explanations (Table I).

TABLE I. COMPARISON WITH THE CLOSEST PRIOR WORK

Study	Dataset	How "concept" is defined	Checked across multiple models?
AiTiALVSD [7]	Short clinical ECG	Diagnostic label, hand-picked examples	No
TCAV-for-EEG [8]	EEG	Hand-picked examples	No
This work	**LTAFDB (continuous, ~24h)**	**Operationalized directly from the dataset's own labels**	**Yes (5 models x 3 depths)**

III. METHODS

Our approach proceeds in two stages. First, we train a standard detection model on raw ECG windows. We then test whether the trained model's predictions are associated with specific clinical concepts, using concept activation vectors (CAVs), and validate any resulting findings with two additional checks: an AF-label confound control and a test for reproducibility across independently trained models.

A. Data and Model

We used the Long Term AF Database [17], comprising 84 continuous ECG recordings of approximately one day each. Each recording was segmented into 10-second windows and labeled AF or non-AF using the database's own rhythm annotations. A compact convolutional neural network (AFNet, Fig. 1) was trained to classify each window, with patients used for testing kept entirely separate from those used for training. The network consists of a stem convolution followed by four residual blocks of increasing channel width, global average pooling, and a classification head, consistent with recent 1D convolutional approaches to inter-patient ECG arrhythmia classification [18]. The trained model distinguished AF from non-AF windows with high accuracy (AUC 0.976 on unseen patients).

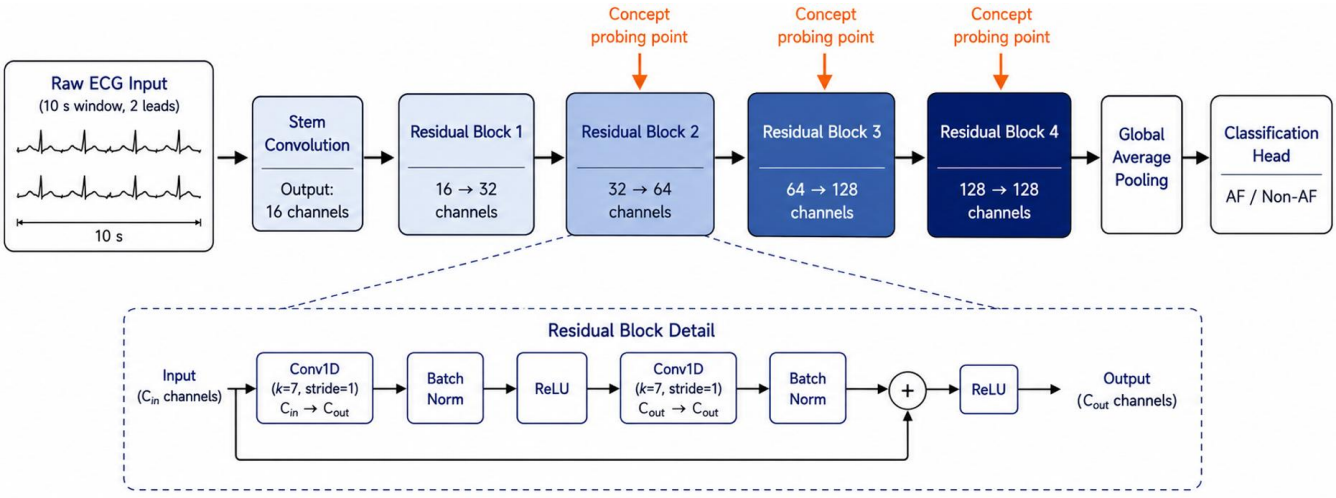

Fig. 1. AFNet architecture. Raw ECG windows pass through a stem convolution and four residual blocks of increasing channel width, followed by global average pooling and a classification head. Concept activation vectors are probed at the output of residual blocks 2, 3, and 4.

B. Concepts and CAV Probing

We defined three clinical concepts directly from the dataset's own beat and rhythm annotations: RR-interval irregularity (the variability in spacing between consecutive heartbeats), ectopic beat burden (the proportion of abnormal beats present), and proximity to a rhythm change (how close a window lies, in time, to a transition between rhythms). For each concept, at three depths within the network, we tested whether the model's internal representation aligned with that concept, using concept activation vectors (CAVs), illustrated in Fig. 2. Briefly: we identify example windows that do and do not exhibit the concept (panel A), find the direction within the model's internal representation that separates them (panel B), and test whether movement in that direction is associated with increased model confidence in an AF prediction (panel C); the proportion of test windows for which this holds is the TCAV score (panel D). Each analysis was repeated across 200 trials: in each trial we drew a fresh random subsample of up to 300 positive and 300 comparison windows, without replacement, from the pooled training and validation set, and trained a new linear CAV (logistic regression) on that subsample; a matched null CAV was trained in parallel on two randomly-labeled subsamples that ignore the concept value. These 200 trials are resampled CAV constructions within the same trained model, dataset, and target set; they are therefore not independent model-training replicates and should not be interpreted as additional model-seed evidence. This within-analysis resampling estimates the variability of the CAV/TCAV comparison conditional on a fixed trained model, whereas the

five model seeds in Section III-C change the learned network itself. The resulting concept-directed and null TCAV score distributions were compared using a Welch's t-test, a Mann-Whitney U test, and a permutation test (10,000 resamples, difference-of-means statistic, label permutation). We treat the three tests as convergent checks on a single comparison rather than independent evidence; the permutation test, being the most assumption-free of the three, was used for the primary significance decision. Benjamini-Hochberg false discovery rate (FDR) correction was applied to the permutation-test p-values across the family of nine concept-depth comparisons, with significance declared at an FDR-adjusted alpha of 0.05.

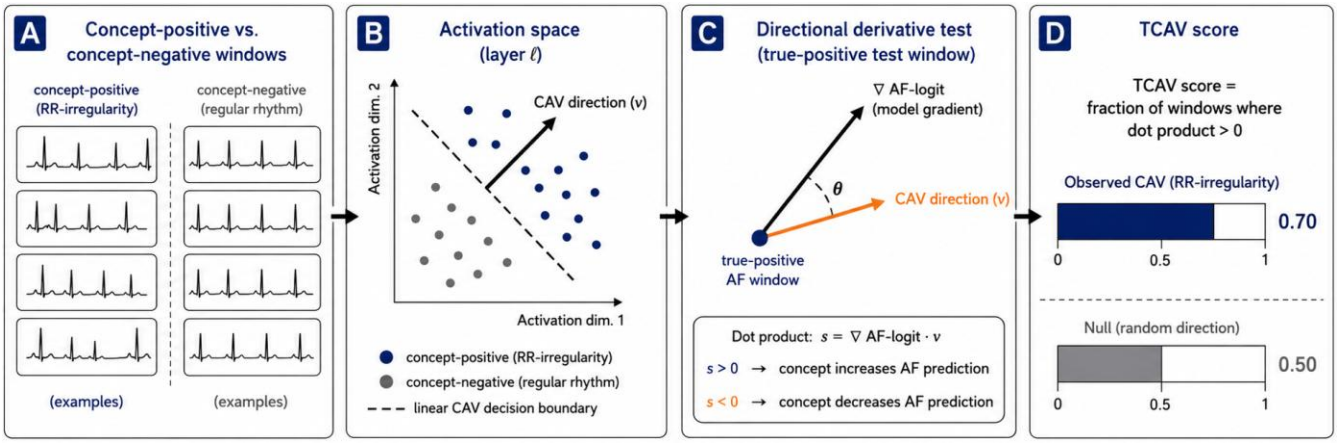

Fig. 2. How a concept activation vector (CAV) is built and tested. (A) Example ECG windows with and without a clinical concept. (B) A linear boundary separating them within the model's internal representation. (C) Testing whether that direction aligns with increased model confidence in AF. (D) The resulting TCAV score.

C. AF-Label Confound Control and Model-Seed Reproducibility

A potential confound arises because these clinical concepts are naturally more prevalent in AF windows: AF windows exhibit greater RR-interval irregularity, for example, simply because irregularity is a defining feature of the rhythm. As a result, a CAV could, in principle, re-derive the AF/non-AF distinction rather than capture an independently represented clinical concept. To address this specific, most direct confound, we reconstructed each CAV using only non-AF-labeled windows (Fig. 3, panels A-B), removing the ability to rely on the AF label directly. The purpose of this AF-label confound control is not to establish that these concepts are entirely independent of AF — RR-irregularity and ectopic burden are, by clinical definition, correlated with the rhythm — but to test whether the CAV's association with the model's AF prediction persists once the direct AF/non-AF label separation is removed from CAV construction. Patient-specific characteristics, recording conditions, and other correlated clinical variables are not controlled for by this step.

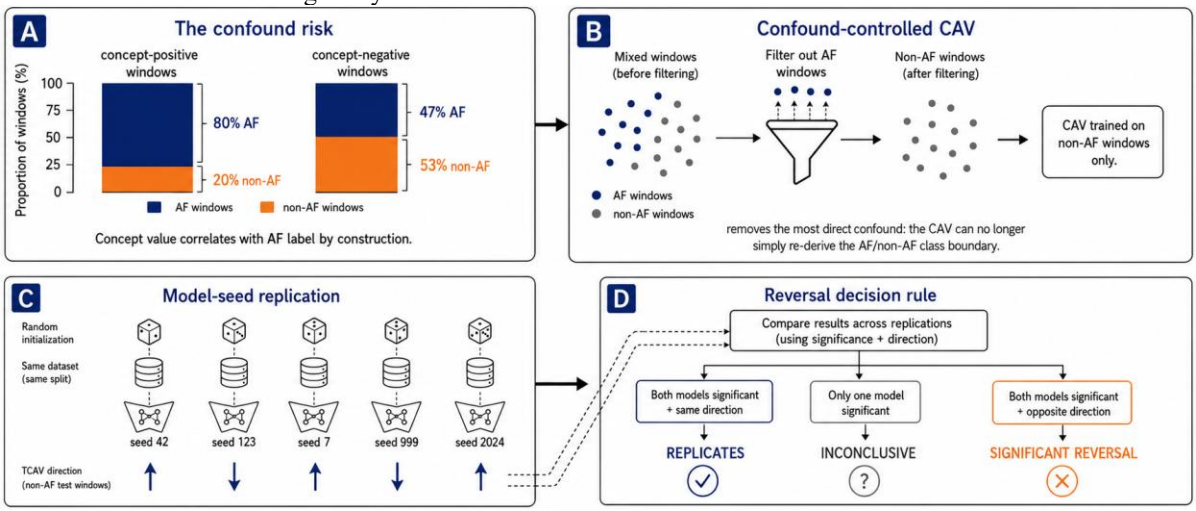

Fig. 3. Two additional validation steps. (A-B) AF-label confound control: rebuilding the CAV from non-AF windows only. (C) Retraining the model five times with different random seeds. (D) The criterion used to determine whether a disagreement between models constitutes a genuine reversal.

A further question is whether a result that survives this AF-label confound control reflects a general property of the model, or an artifact of a single training run. We trained four additional copies of AFNet, each initialized with a different random seed (varying both the model's weight initialization and the order in which training batches were drawn, while holding the data split and all other hyperparameters fixed) and repeated the full confound-controlled analysis on each (Fig. 3,

panel C). We designate the original model (seed 42) as the reference and compare each of the other four models against it individually. A result was classified as a significant reversal only when it was statistically significant in both the reference model and at least one comparison model, but in opposite directions; a result was classified as replicating when it was significant and in the same direction in the reference model and all four comparison models; all other outcomes were classified as inconclusive (Fig. 3, panel D). This criterion is intentionally asymmetric: replication required complete directional consistency across all five models, whereas any single statistically significant opposite-direction result was sufficient to flag potential instability. Because this rule can make the reversal category more sensitive than the replication category, we also performed a simple threshold-sensitivity check using the per-model results in Table III: requiring at least two, rather than one, comparison models to disagree significantly does not change which four concept-depth combinations are labeled reversals; a threshold of three would be more stringent and would retain only the combination with three dissenting models. We therefore interpret the reversal count as a conservative instability flag rather than as an estimate of the prevalence of reversals. The individual model outcomes are reported in Section IV-C so that readers can assess the degree of disagreement directly.

IV. Results

A. The Pattern Depends on How Deep You Look

Before applying any confound control, we tested whether the model's internal representation aligned with each of the three clinical concepts at three depths within the network. Fig. 4 plots the resulting TCAV score against its null baseline for each concept-depth combination.

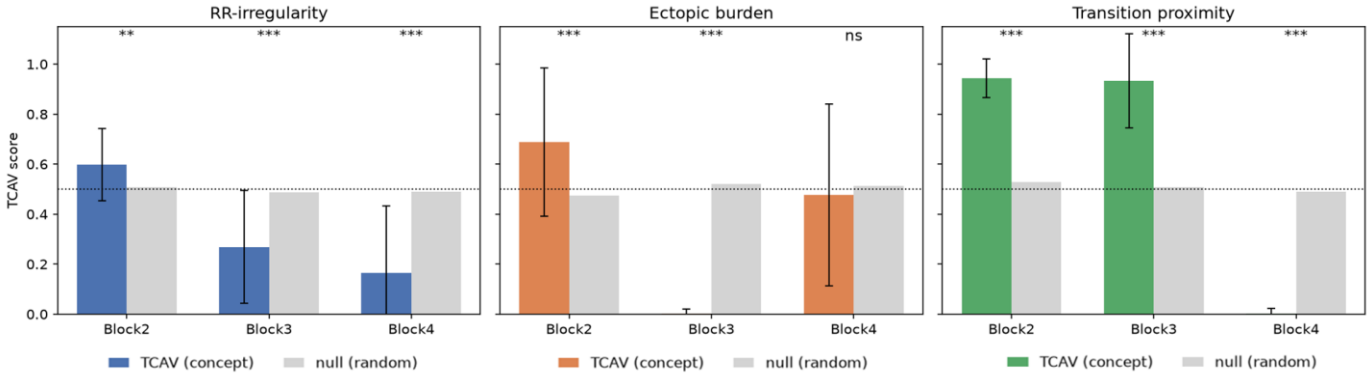

Fig. 4. TCAV score compared to a random-direction baseline, at each depth, for each pattern. Error bars show ±1 SD across 200 resampled trials; asterisks mark significance after FDR correction (* p<0.05, ** p<0.01, *** p<0.001, ns = not significant).

Nearly every result in Fig. 4 is statistically significant; the one exception is ectopic beats at the final layer. Irregular timing and ectopic beats are both positively associated with AF near the start of the network, then reverse to a negative association by the middle layer. The distance-to-rhythm-change pattern shows the opposite and more pronounced behavior: strongly positive through the first two depths, then reversing to strongly negative immediately before the model's final decision.

B. The AF-Label Confound Check Changes Three Results

To test whether these associations reflected the clinical concept itself rather than a proxy for the AF label, we rebuilt every CAV using only non-AF-labeled windows and repeated all nine comparisons. Table II reports the outcome of each comparison, alongside the corresponding uncontrolled result.

TABLE II. TCAV Significance and Direction Before and After AF-Label Confound Control, Across All Nine Concept-Depth Combinations

Concept	Layer	Uncontrolled	AF-Controlled	Change
Irregular heartbeat timing	Early	+ (p<0.01)	+ (p<0.001)	—
Irregular heartbeat timing	Middle	− (p<0.001)	− (p<0.001)	—
Irregular heartbeat timing	Final	− (p<0.001)	+ (p<0.001)	Reversed
Ectopic (abnormal) beats	Early	+ (p<0.001)	+ (p<0.001)	—
Ectopic (abnormal) beats	Middle	− (p<0.001)	− (p<0.001)	—
Ectopic (abnormal) beats	Final	not significant	+ (p<0.001)	Gained significance
Distance to rhythm change	Early	+ (p<0.001)	+ (p<0.001)	—
Distance to rhythm change	Middle	+ (p<0.001)	not significant	Lost significance
Distance to rhythm change	Final	− (p<0.001)	− (p<0.001)	—

As Table II shows, six of the nine comparisons were unaffected by the AF-label confound control. The remaining three changed in three different ways: the irregular-timing result reversed sign at the final layer, the distance-to-rhythm-change result lost significance at the middle layer, and the ectopic-beats result at the final layer gained significance where

none had existed before. These changes indicate that the observed association was sensitive to the AF-label composition of the CAV training set, not necessarily that the original result was purely an artifact, since a change in direction or significance could also reflect a shift in the CAV's measured geometry once the training population changed. We treat all three changes as evidence that the corresponding associations warrant further scrutiny, rather than as proof of which version — uncontrolled or AF-label-controlled — is the more genuine one; Section IV-C subjects all nine associations to the further test of model-seed reproducibility.

C. The Critical Test: Reproducibility Across Retrained Models

We then asked whether a result that survives this AF-label confound control also holds when the model is retrained from scratch under a different random initialization. Fig. 5 summarizes the outcome for all nine concept-depth combinations; Table III reports the underlying per-model result, since the single reversal/replicate/inconclusive label can obscure a more nuanced pattern.

We designate the model trained with seed 42 as the reference and classify each concept-depth combination by comparing the four other models against it individually: a result is a significant reversal only if it is significant in the reference model and in at least one comparison model, with opposite directions; it replicates if it is significant and in the same direction across the reference and all four comparisons; otherwise it is inconclusive. As Table III shows, three combinations replicate under this strict criterion, four show a significant reversal in at least one comparison, and two are inconclusive. The threshold-sensitivity check described in Section III-C shows that requiring two dissenting comparison models instead of one leaves the four reversal labels unchanged, although a threshold of three would produce a more conservative reversal count. These results should therefore be read as evidence of directional sensitivity within the tested AFNet/LTAFDB setting, not as an estimate of a general reversal rate. The individual model outcomes remain important context: for example, distance-to-rhythm-change at the early layer is called a reversal because three of the four comparison models do not all agree with the reference, while the direction tally shows two models significant in the original direction and three significant in the opposite direction. More generally, the results do not show that the model fails to use these concepts; they show that the direction of the concept-prediction association is not fully stable across independently trained instances of the same architecture. Five retrained models demonstrate that this sensitivity to initialization can occur, but they are not enough to establish how frequently it would occur across architectures, datasets, or larger numbers of retraining runs.

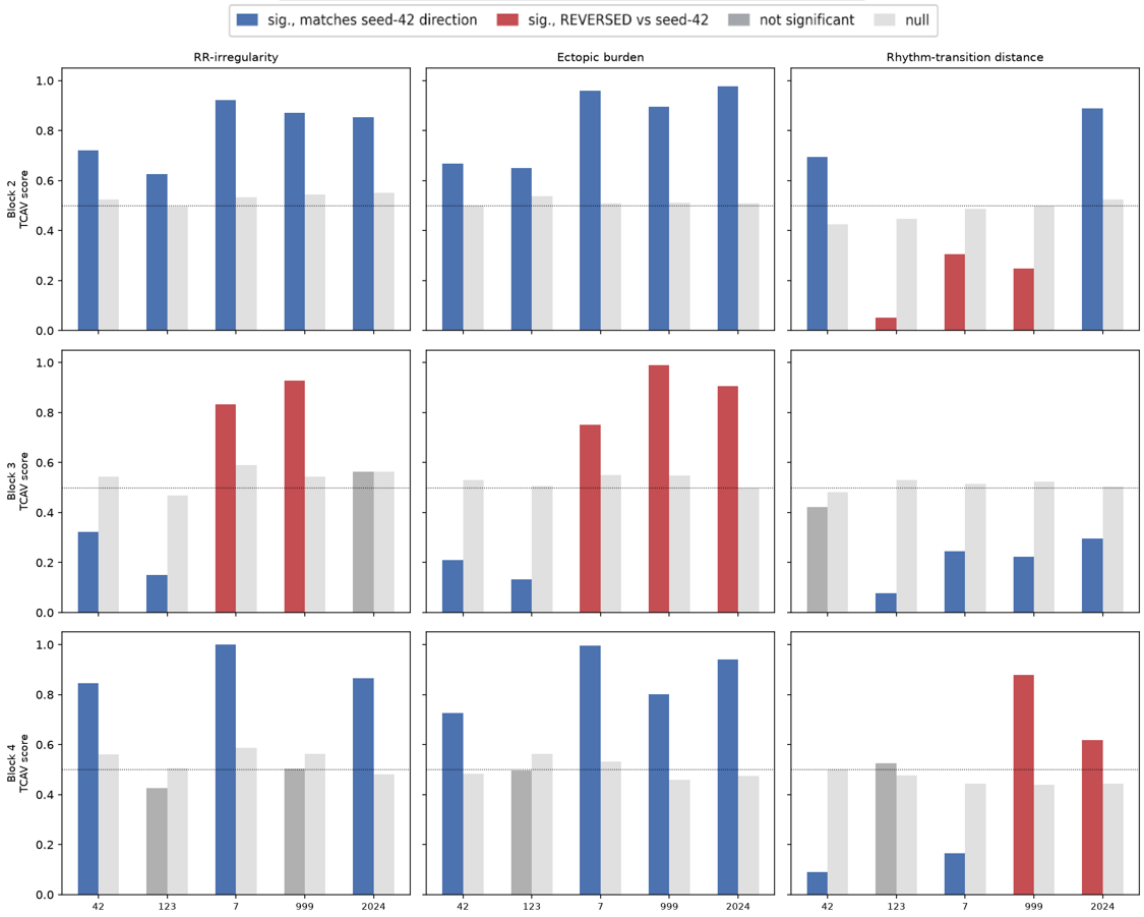

Fig. 5. Results across five independently trained models, for each pattern at each depth. Red bars indicate concept-depth combinations in which at least one comparison model showed a statistically significant association in the direction opposite to the reference model.

TABLE III. DIRECTION AND SIGNIFICANCE OF THE TCAV ASSOCIATION IN EACH OF THE FIVE INDEPENDENTLY TRAINED MODELS. SEED 42 IS THE REFERENCE MODEL

Concept	Layer	42	123	7	999	2024	Direction tally	Classification
Irregular heartbeat timing	Early	+	+	+	+	+	5/5 +, 0/5 −, 0/5 ns	Replicates
Irregular heartbeat timing	Middle	−	−	+	+	*(+)*	2/5 +, 2/5 −, 1/5 ns	Reversal
Irregular heartbeat timing	Final	+	*(−)*	+	*(−)*	+	3/5 +, 0/5 −, 2/5 ns	Inconclusive
Ectopic (abnormal) beats	Early	+	+	+	+	+	5/5 +, 0/5 −, 0/5 ns	Replicates
Ectopic (abnormal) beats	Middle	−	−	+	+	+	3/5 +, 2/5 −, 0/5 ns	Reversal
Ectopic (abnormal) beats	Final	+	*(−)*	+	+	+	4/5 +, 0/5 −, 1/5 ns	Inconclusive
Distance to rhythm change	Early	+	−	−	−	+	2/5 +, 3/5 −, 0/5 ns	Reversal
Distance to rhythm change	Middle	*(−)*	−	−	−	−	0/5 +, 4/5 −, 1/5 ns	Replicates
Distance to rhythm change	Final	−	*(+)*	−	+	+	2/5 +, 2/5 −, 1/5 ns	Reversal

D. Comparison with Standard Attribution Methods

As a point of comparison, we assessed whether two standard sample-level attribution methods, Grad-CAM [19] and SHAP [20] show any relationship with the same two clinical concepts. Table IV summarizes the results. This is a different kind of test than the concept-based analysis above: TCAV asks whether movement in a concept direction within the model's representation is associated with increased AF confidence, whereas this correlation asks whether the raw magnitude of a method's per-window attribution tracks the raw concept value. A near-zero correlation therefore does not mean these methods failed to detect the same effect TCAV detected; it means they were not asked, and did not answer, the same question.

TABLE IV. PEARSON CORRELATION (R) BETWEEN PER-WINDOW ATTRIBUTION MAGNITUDE AND THE RAW CONCEPT VALUE, ON TRUE-POSITIVE AF TEST WINDOWS (GRAD-CAM: N=3,093-3,105; SHAP: N=299-300, SUBSAMPLED FOR COST). * P<0.05; NS = NOT SIGNIFICANT

Method	vs. Irregular Timing	vs. Ectopic Beats
Grad-CAM	r = 0.004 (ns)	r = 0.038 (p = 0.04)*
SHAP	r = 0.012 (ns)	r = 0.053 (ns)

Table IV should therefore be interpreted narrowly. The four reported correlations answer a different question from TCAV: whether the raw magnitude of a sample-level attribution tracks the raw value of a clinical concept. They do not test whether Grad-CAM or SHAP recover the same internal concept direction, depth-dependent association, or sign tested by TCAV. The near-zero correlations indicate little linear association between these particular attribution magnitudes and the selected raw concept values in this experiment; the nominally significant Grad-CAM/ectopic correlation is extremely small. These results cannot establish that Grad-CAM or SHAP failed to detect the clinical patterns identified by TCAV. Rather, they show that the two attribution methods, under this particular correlation-based comparison, were not measuring the same quantity as the TCAV analysis. A method-specific, matched evaluation would be required to make stronger comparative claims.

V. DISCUSSION

Our main finding does not concern any specific clinical pattern. It is that statistical significance and AF-label confound robustness do not guarantee that a concept-prediction association will reproduce across independently trained models: confound robustness and model-seed reproducibility are distinct validation dimensions, and they can disagree in this case study. In our data, the pattern that appeared most convincing after the confound check — distance-to-rhythm-change at the final layer — was also among the associations that showed directional disagreement after retraining. Meanwhile, three associations proved fully consistent under the strict five-model replication criterion. A concept-based explanation should therefore not be considered validated solely because it is statistically significant and survives a confound check; in this setting, testing stability across independently trained model instances provides useful complementary evidence. This interpretation is consistent with broader calls for more rigorous evaluation of interpretability and explanation methods [14], [15] . Importantly, our evidence is limited to the AFNet architecture, the Long Term AF Database, the three operationalized concepts, and the five training seeds used here. We do not claim that directional instability is universal across architectures, datasets, ECG corpora, or concept definitions. Rather, the study demonstrates that such instability can occur and should be checked when the goal is to treat a concept-based association as a stable property of a trained model.

A few limitations are worth noting. Five retrained models are enough to demonstrate that sensitivity to model initialization can occur, but not enough to precisely estimate how often a given association would replicate under further retraining; our results should be read as evidence of sensitivity rather than as a population-level reproducibility rate. We tested only one model architecture on one dataset of 84 continuous recordings. The AF-label confound control addresses the most direct confound in this setting but does not rule out other correlated clinical or recording-specific variables. The 200 CAV trials are repeated resamplings within a fixed model and dataset, not independent training runs, so they should not be

counted as additional evidence of model-level reproducibility. Finally, the Grad-CAM/SHAP comparison uses a different statistical quantity from TCAV and is consequently descriptive rather than a head-to-head test of explanation quality. These limitations define the intended scope of the conclusions and motivate the extensions outlined below.

VI. Future Work

Future work should test whether the observed directional sensitivity persists across alternative ECG architectures, independent ECG corpora, and longer or differently segmented recordings. The framework should also be extended to additional concept types, including morphology, rate-related, and temporal-transition concepts, and to larger numbers of independent training runs so that reproducibility can be estimated rather than only flagged. For the comparison with sample-level attribution methods, future studies should use matched evaluation questions and method-appropriate metrics rather than comparing TCAV directionality with raw attribution–concept correlations. More broadly, systematic XAI reviews emphasize the need for explicit explanation assessment and trustworthy evaluation in healthcare [9], [11], [12]. These extensions would help determine whether the validation framework demonstrated here is specific to the present AFNet/LTAFDB case or generalizes to other physiological time-series models and explanation methods.

VII. Conclusion

A concept-based AI explanation can be statistically significant and survive a targeted AF-label confound control, yet still fail to reproduce directionally when the same architecture is retrained from different random initializations. In this AFNet/LTAFDB case study, confound checks and model-seed reproducibility checks captured different sources of uncertainty, and the latter exposed directional sensitivity that was not apparent from the original model alone. We therefore recommend treating model-seed reproducibility as a complementary validation check when a concept-based association is intended to be interpreted as a stable property of a trained model. We do not claim that this requirement has been established as a universal standard for all concept-based explanation methods from this single study; validation requirements should be adapted to the explanation method, model, data, and intended use.

References

[1] S. Biton *et al.*, "Generalizable and robust deep learning algorithm for atrial fibrillation diagnosis across geography, ages and sexes," *npj Digit. Med.*, vol. 6, no. 1, p. 44, Mar. 2023, doi: 10.1038/s41746-023-00791-1.

[2] S. Mousavi, F. Afghah, and U. R. Acharya, "HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks," *Computers in Biology and Medicine*, vol. 127, p. 104057, Dec. 2020, doi: 10.1016/j.compbiomed.2020.104057.

[3] M. A. Atiea and M. Adel, "Transformer-based Neural Network for Electrocardiogram Classification," *IJACSA*, vol. 13, no. 11, 2022, doi: 10.14569/IJACSA.2022.0131139.

[4] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," presented at the 35th Int. Conf. Machine Learning (ICML), 2018, pp. 2668–2677.

[5] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," in *33rd Conf. Neural Information Processing Systems (NeurIPS)*, 2019, pp. 9273–9282.

[6] C.-K. Yeh, B. Kim, S. O. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," presented at the 34th Conf. Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2020, pp. 20554–20565.

[7] M. S. Lee *et al.*, "Transparent and Robust Artificial Intelligence-Driven Electrocardiogram Model for Left Ventricular Systolic Dysfunction," *Diagnostics*, vol. 15, no. 15, p. 1837, Jul. 2025, doi: 10.3390/diagnostics15151837.

[8] A. Brenner *et al.*, "Concept-based AI interpretability in physiological time-series data: Example of abnormality detection in electroencephalography," *Computer Methods and Programs in Biomedicine*, vol. 257, p. 108448, Dec. 2024, doi: 10.1016/j.cmpb.2024.108448.

[9] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[10] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity Checks for Saliency Metrics," *AAAI*, vol. 34, no. 04, pp. 6021–6029, Apr. 2020, doi: 10.1609/aaai.v34i04.6064.

[11] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.

[12] S. A. Dubey and A. A. Pandit, "A Comprehensive Review and Application of Interpretable Deep Learning Model for ADR Prediction," *IJACSA*, vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130924.

[13] F. Di Martino and F. Delmastro, "Explainable AI for clinical and remote health applications: a survey on tabular and time series data," *Artif Intell Rev*, vol. 56, no. 6, pp. 5261–5315, Jun. 2023, doi: 10.1007/s10462-022-10304-3.

[14] K. Agrawal, R. El Shawi, and N. Ahmed, "XAI-Eval: A framework for comparative evaluation of explanation methods in healthcare," *DIGITAL HEALTH*, vol. 11, p. 20552076251368045, May 2025, doi: 10.1177/20552076251368045.

[15] G. Manimaran *et al.*, "Explainable deep learning-based techniques for ECG-Based heart disease classification: A systematic literature review and future direction," *Computers in Biology and Medicine*, vol. 199, p. 111324, Dec. 2025, doi: 10.1016/j.compbiomed.2025.111324.

[16] J.-H. Jang *et al.*, "A novel XAI framework for explainable AI-ECG using generative counterfactual XAI (GCX)," *Sci Rep*, vol. 15, no. 1, p. 23608, Jul. 2025, doi: 10.1038/s41598-025-08080-5.

[17] S. Petrutiu, A. V. Sahakian, and S. Swiryn, "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans," *EP Europace*, vol. 9, no. 7, pp. 466–470, Jul. 2007, doi: 10.1093/europace/eum096.

[18] M. E. A. Bourkha, A. Hatim, D. Nasir, S. E. Beid, and A. S. Tahiri, "A Novel Inter-Patient ECG Arrhythmia Classification Approach with Deep Feature Extraction and 1D Convolutional Neural Network," *IJACSA*, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150265.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.

[20] V. Kumar, R. K. Burman, A. Kumar, N. Kumar, Md. S. Alam, and P. Kumar, "Explainable Artificial Intelligence(XAI) in Cardiac Disease: Using SHAP Technique," in *Recent Trends in AI Enabled Technologies*, vol. 2322, G. Paidi, A. K. Varma, S. R. Kadiri, and T. K. R. Bollu, Eds., in Communications in Computer and Information Science, vol. 2322. , Cham: Springer Nature Switzerland, 2026, pp. 105–118. doi: 10.1007/978-3-031-92041-7_9.